Relightable Neural Human Assets from Multi-view Gradient Illuminations

Taotao Zhou[1,4*] Kai He[1*] Di Wu[1,3*] Teng Xu[1,4] Qixuan Zhang[1,3]
Kuixiang Shao[1,4] Wenzheng Chen[2] Lan Xu[1†] Jingyi Yu[1†]
[1]ShanghaiTech University [2]University of Toronto [3]Deemos Technology
[4]LumiAni Technology
{zhoutt, hekai, wudi, xuteng, zhangqx1, shaokx, xulan1, yujingyi}@shanghaitech.edu.cn
wenzheng@cs.toronto.edu

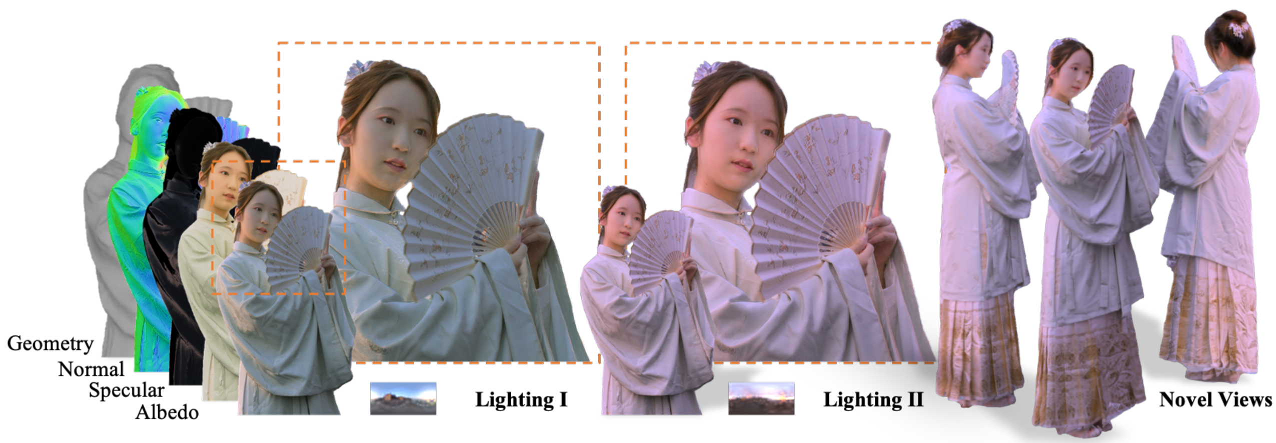

Figure 1. We present UltraStage, a new dataset containing more than 2000 human assets captured under multi-view and multi-illumination settings. The high-quality images allow us to extract detailed normal, albedo, and material maps, as well as reconstruct fine geometry (left). We further propose a neural 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 [1,29]. Earlier MVS typically exploits feature matching and bundle adjustment to find ray correspondences

*Equal contribution.
†Corresponding author.

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 the public due to challenges on multiple fronts for constructing such imaging systems.

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.

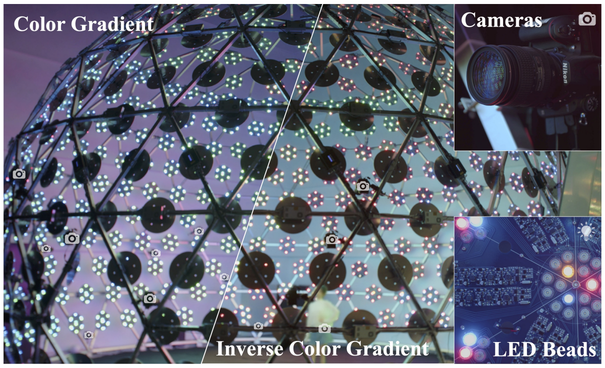

Figure 3. System Overview. The PlenOptic Stage Ultra is an 8-meter lighting stage composed of 32 sounding cameras and 22,080 controllable light sources. It supports both MVS and PS capture settings, enabling high-quality geometry and material acquisition for large-scale subjects and objects. Note that a human is inside in the middle of the picture.

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 recovering 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-

Hardware	#Camera	#Light	Radius	Resolution	Color light
Wenger *et al*. [87]	1	468	1m	1K	-
Kampouris *et al*. [35]	1	336	1.25m	-	✓
Guo *et al*. [27]	90	21k	-	4K	✓
Peng *et al*. [67]	21	-	-	-	-
Bi *et al*. [2]	**140**	460	1.1m	4K	-
Ours	32	**22k**	**4m**	**8K**	✓

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.

ing in various activities, such as walking, running, standing, stretching, working out, and taking selfies. The dataset also encompasses diverse clothing styles, including classical attire, sportswear, and casual outfits. We captured scenes of individuals interacting with objects, as well as multi-person scenarios.

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:

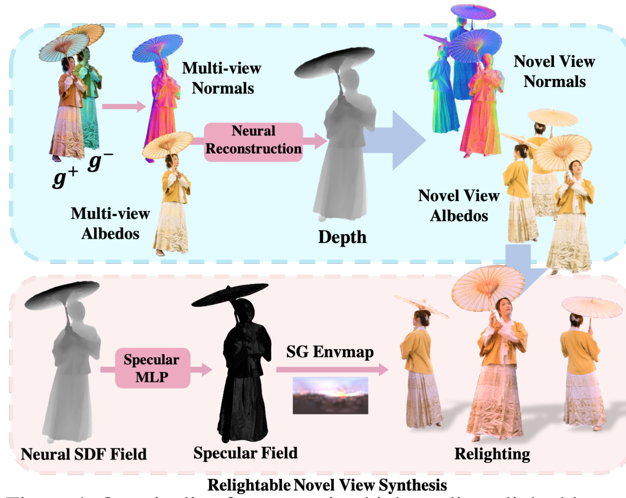

Figure 4. Our pipeline for generating high-quality relightable neural human assets. Given the multi-view input, we first extract corresponding PS normal maps and then train a Neural SDF field from these normal maps (top left). We further adopt depth-guided neural texturing to synthesize novel-view normal and albedo maps (top right) and optimize a neural material field (bottom left) for photo-realistic relighting tasks(bottom, right).

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

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

Dataset	#Frm	#Subj	#View	Res	Relightable	Normal
3DPW [84]	51k	7	1	-	-	-
AIST++ [49]	10M	30	9	-	-	-
HuMMan [5]	**60M**	**1k**	12	1K/4K	-	-
ICT-3DRFE [76]	∼14k	23	2	1K	✓	✓
Dynamic OLAT [95]	603k	36	1	4K	✓	✓
UltraStage(Ours)	192k	100	**32**	**8K**	✓	✓

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).

surface albedo $\mathbf{a}$ to be g^0, details are included in supplementary materials. We do this with two observations: 1) White light helps produce minimum shadows on the human body, largely reducing the influence of self-occlusions. 2) In gradient illuminations normal and albedo are entangled, whereas images captured in white light preserve most albedo information. Although shading is still backed in the albedo, practically, we find it works quite well in all our experiments. Please see Fig. 1 and Fig. 2 for examples. For other surface reflectance parameter estimation, we solve it with neural representation (Sec. 4).

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 2 compares UltraStage Dataset and most existing human 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 Tab. 2. Existing publicly available relightable datasets are all face datasets, such as ICT-3DRFE [76] with two cameras and single-view Dynamic OLAT [95], with one or two cameras. Human datasets like [5, 49, 84] contain large-scale frames but do not support relighting. UltraStage is the first relightable human-centric dataset with 32 viewpoints and the highest resolution.

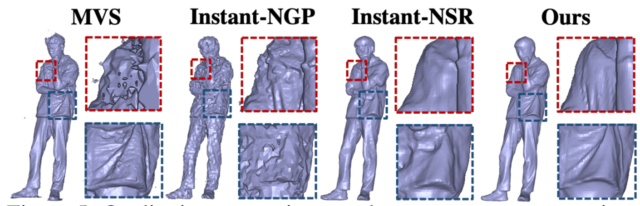

Figure 5. Qualitative comparison on the geometry reconstruction. We compare our method with traditional MVS solution, Instant-NGP [59], Instant-NSR [99]. Our method applies PS normal maps to guide geometry generation, demonstrating detailed reconstruction effects.

4. Neural Asset Generation

UltraStage presents high-quality normal and albedo maps under fixed viewpoints. Inspired by recent advances in neural representations, we further model each example with neural networks, turning it into a neural human asset that enables relightable novel view synthesis.

As shown in Fig. 4, our neural processing pipeline consists of two stages. In the first stage, we take the high-quality normal maps as guidance, training a signed distance field (SDF) [85, 99] to represent human geometry. We demonstrate in Fig. 5 how normal priors significantly improve geometry quality, *e.g.*, faithfully reconstructing details such as folds and wrinkles on clothing.

In the second stage, we adopt deferred rendering techniques that can effectively synthesize free-view renderings of the subject under new and more sophisticated lighting environments. We first prepare a set of G-buffers, including normal, albedo, and material maps. To fully leverage our high-quality PS normal and albedo maps, we devise a depth-guided texture blending method, akin to [99], to synthesize more detailed albedo and normal buffers and apply inverse rendering frameworks [98] to generate material buffer. Once completed, we can shade the image with any desired lighting, which we elaborate on in Sec. 4.2.

4.1. High-quality Neural Geometry Modeling

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: $F_{SDF} : \mathbf{x} \mapsto s$ that maps each 3D location $\mathbf{x} \in \mathbb{R}^3$ to its Signed Distance (SD) value $s \in \mathbb{R}$.

However, instead of supervising with RGB images, we

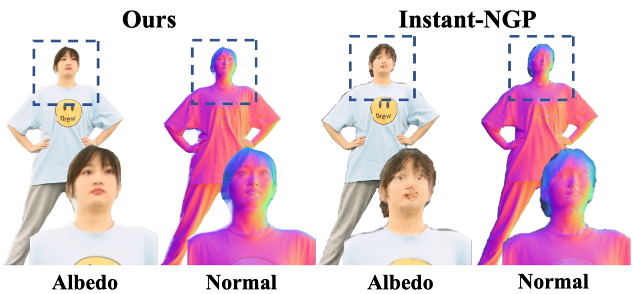

Figure 6. Qualitative comparison on depth-guided G-buffer generation. We compare our method with volume rendering techniques. Specifically, we train normal and albedo neural fields with Instant-NGP [59] and apply volume rendering to generate corresponding G-buffers. By explicitly employing the high-quality 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 $\mathbf{c} \in SE(3)$, we first generate corresponding G-buffers, including the normal map $\mathbf{N} \in \mathbb{R}^{h\times w\times 3}$, albedo map $\mathcal{A} \in \mathbb{R}^{h\times w\times 3}$, material map $\mathcal{M} \in \mathbb{R}^{h\times w\times 1}$. To take advantage of the photometric priors, we utilize the pretrained SDF field to synthesize a depth buffer $\mathcal{D} \in \mathbb{R}^{h\times w\times 1}$, 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 $\mathbf{x}$ from $\mathcal{D}$, normal $\mathbf{n}$ from $\mathbf{N}$, albedo $\mathbf{a}$ from $\mathcal{A}$, and material $\mathbf{m}$ from $\mathcal{M}$. The outgoing radiance L_o in viewing direction w_o is then computed as:

[1] AgiSoft PhotoScan Professional (Version 1.8.4) (Software). (2022*). Retrieved from http://www.agisoft.com/downloads/installer/

	Albedo Map			Normal Map		
Method	PSNR↑	SSIM↑	RMSE↓	Mean	<5°	<25°
Ours	**30.868**	**0.972**	**0.029**	**4.296°**	**80.848%**	**96.293%**
Instant-NGP	30.529	0.968	0.030	4.314°	79.891%	96.013%

Table 3. Quantitative evaluation on the quality of novel view albedo and normal maps. We separate two views from captured views for testing, and evaluate albedo loss and normal vector error in angular. Our reprojected albedo and normal maps produce higher precision in comparison to volume rendering of Instant-NGP [59].

$$L_o(w_o;\mathbf{x}) = \int_\Omega L_i(w_i;\mathbf{x}) f_r(w_o, w_i;\mathbf{x})\,|w_i \cdot n|\,dw_i \ , \tag{2}$$

where $L_i(w_i;x)$ is the incident radiance at position $\mathbf{x}$ from direction w_i , $f_r(w_o, w_i; x)$ is the BRDF that consumes material parameters $\mathbf{a}$ and $\mathbf{m}$, 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 $\mathbf{x}$. We then following NeuS [85] to convert it into density δ:

$$\delta(\mathbf{x}) = \max\left(\frac{-\frac{\mathrm{d}\Phi_s}{\mathrm{d}t}(f_{\text{sdf}}(\mathbf{x}))}{\Phi_s(f_{\text{sdf}}(\mathbf{x}))}, 0\right) \ , \tag{3}$$

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

$$\mathcal{D}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\delta(\mathbf{r}(t))t dt, \tag{4}$$

where $T(t) = \exp\left(-\int_{t_n}^{t} \delta(\mathbf{r}(s))ds\right)$ 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 $\mathbf{x}$ at each pixel. We then adopt a neural depth-guided reprojection technique [17, 99] to decide its normal $\mathbf{n}$ and albedo $\mathbf{a}$. Specifically, we reproject $\mathbf{x}$ into nearby K PS views to query the corresponding high-quality normal and albedo values $\{\mathbf{n}_i, \mathbf{a}_i\}_{i=1}^K$. The normal and

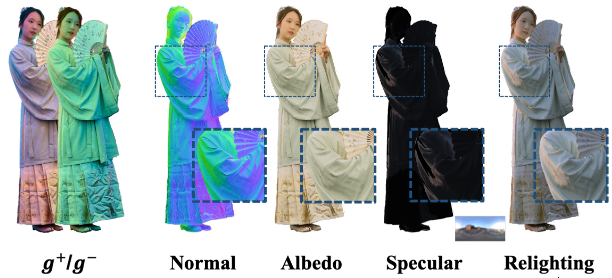

Figure 7. Given color gradient illumination observations g^+, g^-, by taking high-quality normal and albedo maps as priors, we learn a neural material field on the pre-trained neural geometry surface, enabling photo-realistic relighting results. Note that the specular map is rendered in the target illumination.

albedo in the novel view are computed by the weighted blending of all the queried candidates, where $(\mathbf{n}, \mathbf{a}) = (\sum_{k=1}^{K} w_k \mathbf{n}_k, \sum_{k=1}^{K} w_k \mathbf{n}_k)$.

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 $\mathbf{x}$ 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: $F_{MAT.} : \mathbf{x} \mapsto r$ that maps each 3D location $\mathbf{x} \in \mathbb{R}^3$ 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.

[2] https://web.twindom.com/

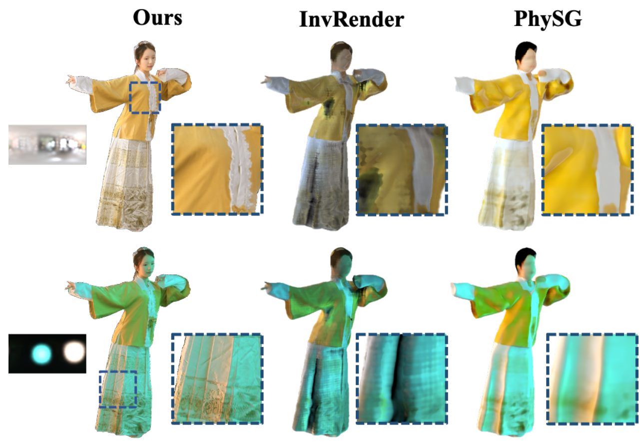

Figure 8. Qualitative comparison on relighting under novel viewpoint with two recent neural relightable novel view synthesis approaches [93, 98]. Our method takes advantage of photometric priors enabling realistic rendering under complex novel illuminations, with high-frequency details preserved.

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

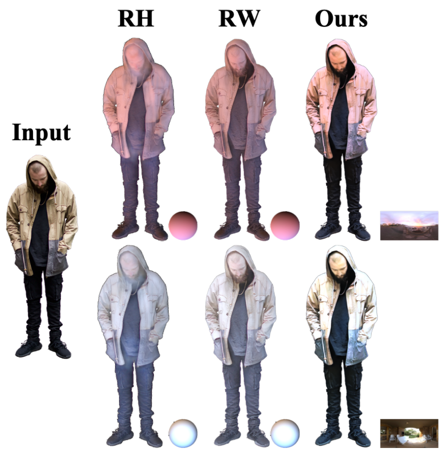

Figure 9. Our relighting results produce sharper details, especially on clothes textures. The reflectance under novel illuminations is more realistic.

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 hours with a learning rate of 10^{-5}. After training, we then apply networks to arbitrary human images and predict the corresponding normal and albedo maps. Although albedo and normal only admit diffuse relighting effects, we find that the results are visually appealing, thanks to the powerful training dataset. We believe our dataset can facilitate a series of relighting research. We will also make this relighting dataset public.

[3]https://hdrihaven.com/

	% Ours is preferred
vs RH [36]	96.6 %
vs RW [80]	83.3 %

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

Experiments. We compare our model with two open-source human relighting methods RW [80] and RH [36], with the former being a following-up work of the latter. We qualitatively compare our relighting results with them in Fig. 9, where the image is selected from the testing dataset in [36]. Our model produces more detailed albedo and accurate normal maps, resulting in superior relighting effects, as evidenced by the distinct cloth wrinkles and natural shading. It is worth noting that we achieve these outcomes solely through basic normal and albedo prediction networks. Applying more advanced designs [67] will definitely improve the effects, which we leave for future works.

We further conduct a human study in Tab. 4 on all the testing images in [36]. It contains 6 testing images in total. We relight each one with 5 new illuminations, resulting in 30 examples. Participants are asked to select the more natural-looking relighting, and our approach receives more than 80% preferences, further attesting to the efficacy of our dataset. More details of the human study can be found in the supplementary materials.

6. Conclusions

We have presented UltraStage, a novel 3D human dataset to bridge human modeling and relighting, which contains more than 2,000 high-quality human assets captured under multi-view and multi-illumination settings unseen before. We further interpret this rich captured input into neural human assets, allowing photo-realistic novel-view synthesis under arbitrary lighting conditions. Extensive experiments demonstrate the superiority of UltraStage. With the above unique characteristics and rareness of our UltraStage, we believe it is critical for future research about high-quality human modeling, with endless potential applications.

Acknowledgements. We thank Hongyang Lin, Qiwei Qiu and Qingcheng Zhao for building the hardware. This work was supported by National Key R&D Program of China (2022YFF0902301), NSFC programs (61976138, 61977047), STCSM (2015F0203-000-06), and SHMEC (2019-01-07-00-01-E00003). We also acknowledge support from Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

References

[1] Ronen Basri, David Jacobs, and Ira Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of computer vision*, 72(3):239–257, 2007. 1, 2

[2] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 2, 4

[3] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 3

[4] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *Advances in Neural Information Processing Systems*, 34:10691–10704, 2021. 3

[5] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022. 3, 5

[6] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008. 2

[7] Xuan Cao, Zhang Chen, Anpei Chen, Xin Chen, Shiying Li, and Jingyi Yu. Sparse photometric 3d face reconstruction guided by morphable models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4635–4644, 2018. 3

[8] Charles-Félix Chabert, Per Einarsson, Andrew Jones, Bruce Lamond, Wan-Chun Ma, Sebastian Sylwan, Tim Hawkins, and Paul Debevec. Relighting human locomotion with flowed reflectance fields. In *ACM SIGGRAPH 2006 Sketches*, pages 76–es. 2006. 8

[9] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9429–9439, 2019. 3

[10] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 3

[11] Wenzheng Chen, Joey Litalien, Jun Gao, Zian Wang, Clement Fuji Tsang, Sameh Khamis, Or Litany, and Sanja Fidler. Dib-r++: Learning to predict lighting and material with a hybrid differentiable renderer. *Advances in Neural Information Processing Systems*, 34:22834–22848, 2021. 2, 3, 6, 7

[12] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4), jul 2015. 3

[13] Ingemar J. Cox, Sunita L. Hingorani, Satish B. Rao, and Bruce M. Maggs. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3):542–567, 1996. 2

[14] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 3

[15] Paul Debevec, Andreas Wenger, Chris Tchou, Andrew Gardner, Jamie Waese, and Tim Hawkins. A lighting reproduction approach to live-action compositing. *ACM Transactions on Graphics (TOG)*, 21(3):547–556, 2002. 2

[16] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. Fusion4d: Real-time performance capture of challenging scenes. *ACM Trans. Graph.*, 35(4), jul 2016. 3

[17] Martin Eisemann, Bert De Decker, Marcus Magnor, Philippe Bekaert, Edilson De Aguiar, Naveed Ahmed, Christian Theobalt, and Anita Sellent. Floating textures. In *Computer graphics forum*, volume 27, pages 409–418. Wiley Online Library, 2008. 6

[18] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 2

[19] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 2

[20] Graham Fyffe. Cosine lobe based relighting from gradient illumination photographs. In *SIGGRAPH'09: Posters*, pages 1–1. 2009. 2, 4

[21] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Trans. Graph.*, 36(6), nov 2017. 2, 8

[22] Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Tinne Tuytelaars, and Luc Van Gool. What is around the camera? In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5180–5188, 2017. 2

[23] Abhijeet Ghosh, Tongbo Chen, Pieter Peers, Cyrus A Wilson, and Paul Debevec. Estimating specular roughness and anisotropy from second order spherical gradient illumination. In *Computer Graphics Forum*, volume 28, pages 1161–1170. Wiley Online Library, 2009. 2

[24] Abhijeet Ghosh, Tongbo Chen, Pieter Peers, Cyrus A Wilson, and Paul Debevec. Circularly polarized spherical illumination reflectometry. In *ACM SIGGRAPH Asia 2010 papers*, pages 1–12. 2010. 2

[25] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–10, 2011. 2, 3

[26] M. Goesele, B. Curless, and S.M. Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2402–2409, 2006. 3

[27] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 2, 3, 4, 8

[28] Oshri Halimi, Fabian Prada, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, and Yaser Sheikh. Garment avatars: Realistic cloth driving using pattern registration. *arXiv preprint arXiv:2206.03373*, 2022. 3

[29] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 3

[30] Tim Hawkins, Jonathan Cohen, and Paul Debevec. A photometric approach to digitizing cultural artifacts. In *Proceedings of the 2001 conference on Virtual reality, archeology, and cultural heritage*, pages 333–342, 2001. 2

[31] Berthold KP Horn and Michael J Brooks. The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33(2):174–208, 1986. 3

[32] Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. Geometry-aware single-image full-body human relighting. In *European Conference on Computer Vision*, pages 388–405. Springer, 2022. 3, 8

[33] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 6

[34] Christos Kampouris and Abhijeet Ghosh. Icl multispectral light stage: building a versatile led sphere with off-the-shelf components. In *Proceedings of the Eurographics 2018 Workshop on Material Appearance Modeling*, pages 1–4, 2018. 2

[35] Christos Kampouris, Stefanos Zafeiriou, and Abhijeet Ghosh. Diffuse-specular separation using binary spherical gradient illumination. In *EGSR (EI&I)*, pages 1–10, 2018. 2, 4

[36] Yoshihiro Kanamori and Yuki Endo. Relighting humans: Occlusion-aware inverse rendering for full-body human images. *ACM Trans. Graph.*, 37(6), dec 2018. 2, 3, 8

[37] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2

[38] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[39] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[40] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncertainty-aware deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12601–12611, 2022. 3

[41] Jaewon Kim, Shahram Izadi, and Abhijeet Ghosh. Single-shot layered reflectance separation using a polarized light field camera. 2016. 2

[42] Martin Klaudiny and Adrian Hilton. High-detail 3d capture and non-sequential alignment of facial performance. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 17–24. IEEE, 2012. 3

[43] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *European conference on computer vision*, pages 82–96. Springer, 2002. 2

[44] Manuel Lagunas, Xin Sun, Jimei Yang, Ruben Villegas, Jianming Zhang, Zhixin Shu, Belen Masia, and Diego Gutierrez. Single-image full-body human relighting. *arXiv preprint arXiv:2107.07259*, 2021. 3, 8

[45] Jean-François Lalonde and Iain Matthews. Lighting estimation in outdoor image collections. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 131–138. IEEE, 2014. 8

[46] Chloe LeGendre, Xueming Yu, Dai Liu, Jay Busch, Andrew Jones, Sumanta Pattanaik, and Paul Debevec. Practical multispectral lighting reproduction. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 4

[47] Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C Buehler, Otmar Hilliges, and Thabo Beeler. Eyenerf: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 3

[48] Jiaman Li, Zhengfei Kuang, Yajie Zhao, Mingming He, Karl Bladin, and Hao Li. Dynamic facial asset and rig generation from a single scan. *ACM Trans. Graph.*, 39(6), nov 2020. 2

[49] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13381–13392, 2021. 3, 5

[50] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE transactions on visualization and computer graphics*, 16(3):407–418, 2009. 2

[51] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 3

[52] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned

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

[53] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. *Rendering Techniques*, 2007(9):10, 2007. 2, 3, 4

[54] Wan-Chun Ma, Andrew Jones, Jen-Yuan Chiang, Tim Hawkins, Sune Frederiksen, Pieter Peers, Marko Vukovic, Ming Ouhyoung, and Paul Debevec. Facial Performance Synthesis using Deformation-Driven Polynomial Displacement Maps. *ACM Transactions on Graphics*, 27(5), 2008. 2

[55] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2, 3

[56] Abhimitra Meka, Maxim Maximov, Michael Zollhöfer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. Lime: Live intrinsic material estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6315–6324, 2018. 2

[57] Abhimitra Meka, Rohit Pandey, Christian Haene, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, et al. Deep relightable textures: volumetric performance capture with neural rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–21, 2020. 2, 3

[58] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3, 6

[59] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2, 3, 5, 6

[60] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 3

[61] Giljoo Nam, Chenglei Wu, Min H Kim, and Yaser Sheikh. Strand-accurate multi-view hair capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 155–164, 2019. 3

[62] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, page 536–543, New York, NY, USA, 2005. Association for Computing Machinery. 2

[63] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 343–352, 2015. 3

[64] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003. 3

[65] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 2, 3

[66] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[67] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 4, 8

[68] Peiran Ren, Yue Dong, Stephen Lin, Xin Tong, and Baining Guo. Image based relighting using neural networks. *ACM Trans. Graph.*, 34(4), jul 2015. 2

[69] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 8

[70] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 3

[71] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Humangan: A generative model of human images. In *2021 International Conference on 3D Vision (3DV)*, pages 258–267. IEEE, 2021. 2

[72] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016. 2

[73] Christopher Schwartz, Ralf Sarlette, Michael Weinmann, and Reinhard Klein. Dome ii: A parallelized btf acquisition system. In *Material Appearance Modeling*, pages 25–31, 2013. 2

[74] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006. 3

[75] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 3

[76] Giota Stratou, Abhijeet Ghosh, Paul Debevec, and Louis-Philippe Morency. Effect of illumination on automatic expression recognition: A novel 3d relightable facial database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 611–618, 2011. 5

[77] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021. 2

[78] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Nelf: Neural light-transport field for portrait view synthesis and relighting. In *Eurographics Symposium on Rendering*, 2021. 3

[79] Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T Barron, and Ravi Ramamoorthi. Light stage super-resolution: continuous high-frequency relighting. *ACM Transactions on Graphics (TOG)*, 39(6):1–12, 2020. 2, 3

[80] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. In *Computer Graphics Forum*, volume 40, pages 205–216. Wiley Online Library, 2021. 2, 3, 8

[81] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012. 2

[82] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999. 2

[83] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 2

[84] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, page 614–631, Berlin, Heidelberg, 2018. Springer-Verlag. 3, 5

[85] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 3, 5, 6

[86] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2

[87] Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics (TOG)*, 24(3):756–764, 2005. 2, 4

[88] Sitao Xiang, Yuming Gu, Pengda Xiang, Mingming He, Koki Nagano, Haiwei Chen, and Hao Li. One-shot identity-preserving portrait reenactment. *ArXiv*, abs/2004.12452, 2020. 2

[89] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 2

[90] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (TOG)*, 37(4):126, 2018. 2

[91] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. Ps-nerf: Neural inverse rendering for multi-view photometric stereo. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 266–284. Springer, 2022. 3

[92] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2

[93] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 2, 3, 6, 7

[94] Longwen Zhang, Chuxiao Zeng, Qixuan Zhang, Hongyang Lin, Ruixiang Cao, Wei Yang, Lan Xu, and Jingyi Yu. Video-driven neural physically-based facial asset for production. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–10, 2022. 3

[95] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 802–812, 2021. 2, 3, 5

[96] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, et al. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG)*, 40(1):1–17, 2021. 2, 3

[97] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 3

[98] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18643–18652, 2022. 2, 3, 5, 7

[99] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, and Jingyi Yu. Human performance modeling and rendering via neural animated mesh. *ACM Trans. Graph.*, 41(6), nov 2022. 2, 5, 6, 7

[100] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 2