This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

Relightable Detailed Human Reconstruction from Sparse Flashlight Images

Jiawei Lu, Tianjia Shao, He Wang, Yong-Liang Yang, Yin Yang, and Kun Zhou, *Fellow, IEEE*

Abstract—We present a lightweight system for reconstructing human geometry and appearance from sparse flashlight images. Our system produces detailed geometry including garment wrinkles and surface reflectance, which are exportable for direct rendering and relighting in traditional graphics pipelines. By capturing multi-view flashlight images using a consumer camera equipped with an co-located LED (e.g., a cell phone), we obtain view-specific shading cues that aid in the determination of surface orientation and help disambiguate between shading and material. To enable the reconstruction of geometry and appearance from sparse-view flashlight images, we integrate a pre-trained model into a differentiable physics-based rendering framework. As the learned image features from synthetic data cannot accurately reflect the shading features on real images, which is crucial for the high-quality reconstruction of geometry details and appearance, we propose to jointly optimize the image feature extractor with two MLPs for SDF and BRDF prediction using the differentiable physics-based rendering. Compared with existing methods for relightable human reconstruction, our system is able to produce high-fidelity 3D human models with more accurate geometry and appearance under the same condition. Our code and data are available at http://github.com/Jarvisss/Relightable_human_recon.

Index Terms—Human reconstruction, human relighting, neural implicit field, sparse view reconstruction.

1 Introduction

The reconstruction of the human body from images has been a longstanding problem in the fields of computer vision and graphics. It has a wide range of applications in the film industry, gaming, and virtual reality, allowing individuals to be displayed in virtual worlds. High-fidelity reconstruction often requires sophisticated hardware, such as commercially available body scanners. Additionally, surface materials can be obtained through traditional photometric stereo strategy [1], such as a light stage with programmable lights and dense synchronized cameras [2], [3], [4]. However, these equipment pieces are often too complex to use and prohibitively expensive, which hinders their widespread adoption in consumer-level applications.

Recently, the rapid advancements in deep learning have led to a growing number of studies employing data-driven approaches for human reconstruction [5], [6], [7], [8], [9], [10]. These methods learn strong priors from human scan data and thus minimize the requirement for sophisticated capture hardware to sparse or single cameras. The recent work [11] demonstrates the novel capability to reconstruct relightable humans from a single image by simultaneously predicting light, albedo, and geometry. However, single-view methods struggle to produce detailed full-body reconstructions. Hence, to balance the complexity of capture and the coverage of views, leveraging sparse views has emerged as a promising and increasingly researched direction. The prevalent approach involves extracting features from input images and subsequently projecting 3D points to each view to sample the features in a pixel-aligned manner. The sampled multi-view features are fused by different strategies, e.g. mean pooling [5], [12], self-attention [13], [14], [15], and visibility [8]. The fused feature is further sent to Multi-Layer Perceptrons (MLPs) to reason the geometry and color. The sparse setting allows the network to model the contribution of each view to the final reconstruction, but there is no guarantee that the raw output produced by the trained neural network accurately represents the human. It is mainly because the networks are trained with the supervision from synthetic data (i.e., modeled or scanned human data rendered with different light conditions [16], [17]). Due to the inherent domain gap between the synthetic data and real data, the results of these learning-based methods could inevitably contain non-neglectable artifacts such as broken limbs, or incorrect colors when the input real images are far from the training data distribution (see the first column in Fig. 3 for example).

Another active area of research on 3D reconstruction involves performing per-scene optimization and modeling each scene by neural implicit fields. Popular optimization-based approaches include volume-based rendering of radiance fields [15], [18], [19], [20] and surface-based rendering of signed distance fields [21], [22]. The methods mentioned have shown considerable success in view synthesis and shape reconstruction. However, the lighting and shadow effects are baked into the reconstruction result, limiting their realism in novel environments. Several methods [23], [24], [25], [26], [27], [28] incorporate physics-based rendering into the radiance field or SDF to decouple the shape, material, and lighting. IRON [29] introduces an optimization scheme

- *Jiawei Lu, Tianjia Shao and Kun Zhou are with the State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China. E-mail: {lujiawei23, tianjiashao}@gmail.com, kunzhou@acm.org*
- *He Wang is with the Department of Computer Science, University College London, Gower Street London, WC1E 6BT United Kingdom. E-mail: h.e.wang@leeds.ac.uk*
- *Yong-Liang Yang is with the Department of Computer Science, University of Bath, United Kingdom. E-mail: y.yang@cs.bath.ac.uk*
- *Yin Yang is with Kahlert School of Computing, University of Utah, USA. E-mail: yangzzzy@gmail.com*

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

with flashlight images as input for improved performance. WildLight [30] further proposes to utilize additional flashlight images to reason the object materials in the wild. However, a common drawback of these relighting methods is their time-consuming nature (e.g. ≈ 10 hours for a single scene optimization), primarily due to the requirement for per-scene optimization on the densely captured images.

In this paper, our aim is to develop a user-friendly system that creates high-quality relightable full-body human assets using consumer devices. The system requires addressing the following challenges. Firstly, to produce realistic relighting effects, we need to recover both the geometry details, such as garment wrinkles and the associate material maps, which is non-trivial. Secondly, the system should be capable of producing results within a user-affordable time (i.e., within 10 minutes), so that ordinary users do not have to wait a long time to get their digital models.

To address these challenges, we establish a system that uses a consumer camera equipped with an LED (e.g., a cellphone) to reconstruct a high-quality relightable human model with decoupled rich geometry details and high-fidelity appearances. The system takes sparse-view flashlight images as input and outputs the geometry and material of the subject. The camera and flashlight can be safely treated as co-located as in [29], [31], [32], so the flashlight as a single point light source provides view-specific shading cues that aid in the determination of surface orientation and help disambiguate between shading and material. To enable the disentanglement and reconstruction of geometry and appearance from sparse flashlight images, we propose a hybrid method that integrates the learning-based method and optimization-based method into a unified framework. The key observation is, though learning-based methods have the domain-gap problem yielding artifacts, these artifacts are generally local, like broken limbs or incorrect details. These learning-based methods (e.g. PIFu [5]) have a good capacity to capture globally correct body shapes and poses. By providing learning-based results as a good starting point for the optimization, the difficulty and complexity of optimization are largely reduced, as the optimization only needs to focus on rectifying the errors of local parts and details. Specifically, instead of directly applying a fixed pre-trained model to supply image features to a volume-rendering module for optimizing the geometry and color [15], we propose to jointly optimize the pre-trained image feature extractor with two MLPs for SDF and BRDF prediction using the differentiable physics-based rendering. This is because the reconstruction of geometry detail and appearance is highly sensitive to the captured shadings in flashlight images. The image features from the pre-trained model using synthetic data have a domain gap to the real image features, which cannot accurately reflect the shading features on the real captured image, which will degrade the reconstruction quality (see Fig. 11 for example). In contrast, the joint optimization can effectively adapt the image features to reflect the captured real shadings, and more accurately predict the SDF and BRDF.

With the flashlight setup and joint optimization framework, our system can reconstruct high-quality relightable human models from sparse-view flashlight images. With the high-fidelity shape and material, the system achieves realistic relighting with versatile lighting effects. As only sparse-view images are required, the computational cost is largely reduced, making the system able to produce the results within 10 minutes on average. We conduct extensive experiments on public synthetic datasets and real captured data. Both qualitative and quantitative results show that our system consistently outperforms existing methods for relightable human reconstruction under the same condition.

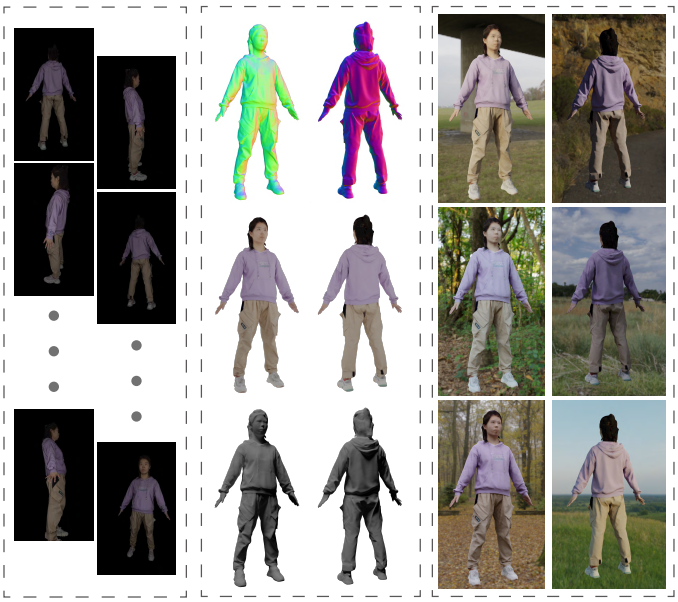


Fig. 1. Given sparse view images of a person captured by a camera with a flashlight (left), our method reconstructs detailed geometry and materials (middle) and allows for the export of textures and relighting. The right two columns depict renderings produced by relighting the reconstruction using different HDR environment light maps from [33].

2 Related Work

2.1 Detailed Human Reconstruction

Reconstructing the detailed 3D shape of a clothed human from single or multi-view images has drawn significant attention in research. Early works are mostly optimization-based. Those methods often get complete 3D shape by using videos of moving subjects [34] or incorporating auxiliary information such as depth [35]. With the rapid progress of deep learning, learning-based methods [5], [6], [9], [10] have emerged for human digitization. These methods achieve detailed clothed human digitization by leveraging pixel-aligned features to predict an implicit occupancy field from either a single image or multiple RGB images. Subsequently, the marching cube algorithm [36] is employed for geometry extraction, and vertex colors can be obtained by a color MLP. However, significant degradation in reconstruction quality is observed when there are substantial deviations in body poses, camera views, and lighting conditions from the training distribution. A set of works [7], [10] attempt to address the problem caused by out-of-distribution poses by incorporating a human parametric model, e.g. SMPL [37]. However, the use of SMPL itself introduces biases due to the misalignment between the parametric model and the clothed human body. As a result, it encounters difficulties when modeling loose clothing and non-body objects. Some

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

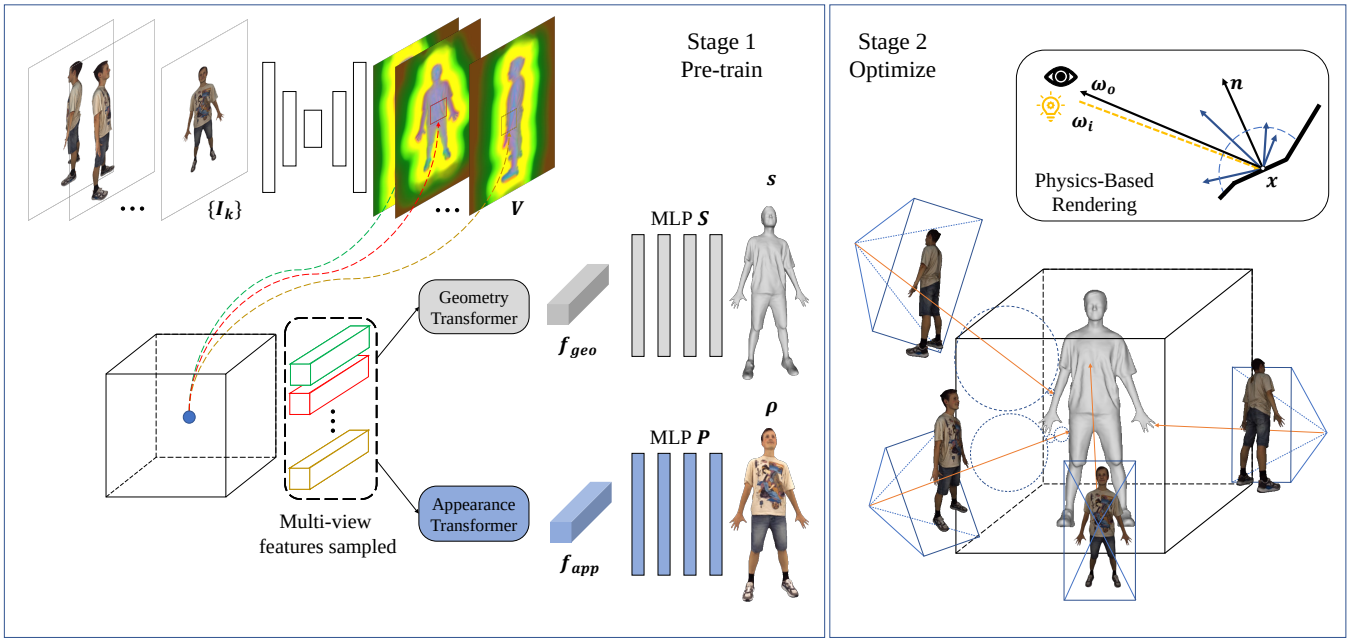


Fig. 2. Our system has two stages. The left part depicts the pre-training stage, points sampled over space and near the surface are projected to each image plane to sample the feature volume in a pixel-aligned fashion. Then, the multi-view features are fused by two transformers into two feature vectors, which are further sent to two MLPs to predict the SDF and BRDF parameters. With the globally correct body shape from the pre-trained model and multi-view flash image observations, we design a joint optimization stage based on sphere tracing and physics-based rendering to further optimize the whole model. This stage effectively rectifies the local errors caused by the domain gap between training and testing data and results in detailed mesh and material reconstruction.

other works rely on extra input such as depth maps [8] or stereo image pairs [38] to further improve the robustness and enhance the details. Econ [39] presents a hybrid method for detailed human recovery from a single view. This method initiates the shape with the SMPL body model and subsequently refines the surface using predicted normal maps, where the reconstruction accuracy relies on the estimation of SMPL and normals. DoubleField [15] proposes the integration of deep learning and volume rendering-based optimization to leverage their respective advantages. However, their method requires sampling a set of points around the surface along each ray for volume rendering. This operation causes decreased performance and large memory consumption, which leads to a limited resolution of feature volume, which prevents them from acquiring high-quality geometry details. In addition, their approach does not support relighting as it lacks material reconstruction capabilities.

2.2 Human Relighting

Relighting of humans has diverse applications in various domains including virtual reality, augmented reality, and digital entertainment. Research mainly focuses on several scenarios, including 1) single image with static unknown lighting condition [40], [41], [42], [43], [44], 2) monocular video-based relighting with static unknown lighting condition, with a static camera and a person in front of the camera and perform dynamic poses [45], [46], [47], and 3) multi-view images with calibrated lights and various lighting patterns, e.g. light-stage based methods [2], [3]. The single-image methods manage to relight a person in an image by learning human-specific priors from large-scale datasets, but are limited to relighting the input image. Video-based approaches mainly optimize a canonical space of the human and enable rendering in novel view, pose, and lighting. However, the reconstruction results often fail to capture fine details due to the highly ill-posed setting, such as the deformed geometry in the video and unknown lighting, resulting in over-smoothed reconstructions and loss of realism [45], [46]. Different from the previous research, our work takes a distinct way of using flashlight images captured by a hand-held device (e.g. a cellphone) for static human reconstruction and relighting. Our setting could achieve free view rendering and produce more reliable relighting results compared to the video-based methods while keeping user-friendliness.

3 Method

3.1 Overview

Aiming at a user-friendly system for human geometry and material reconstruction, we introduce an approach that uses a consumer camera equipped with a flash to reconstruct a high-quality relightable human model. Taking sparse-view flashlight images $\{I_k\}$ of a person captured in a darkroom as input, our model G produces the geometry as signed distance $d \in \mathbb{R}$ and material parameters $\rho \in \mathbb{R}^3, \gamma \in \mathbb{R}, \kappa \in \mathbb{R}^3$ at an arbitrary point x.

$$d, \{\rho, \gamma, \kappa\} = G(\{I_k\}, x; \Theta), \tag{1}$$

where ρ, γ, κ denotes diffuse albedo, roughness, and specular albedo, respectively.

To facilitate the disentanglement and reconstruction of geometry and appearance from sparse flashlight images, we

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

first pre-train a model to extract pixel-aligned features on a synthetic rendered human dataset, and then perform a joint optimization of the pre-trained feature extractor and the two MLPs for SDF and BRDF prediction using differentiable physics-based rendering. The pre-trained model could capture globally correct body shapes and poses, which provides a good base for optimization and significantly reduces the complexity of optimization. However, directly applying the features learned from synthetic data to optimize the MLPs on real data could result in degraded reconstruction results (see Fig. 11), because the learned features cannot accurately reflect the shading features on real images. Therefore, we propose to jointly optimize the feature extractor and the MLPs in the optimization stage, producing more accurate reconstruction.

In the following sections, we will delve into the specifics of these two stages.

3.2 Pre-training

We initially pre-train our model using synthetic data that contains ground truth information. The pre-training stage allows the model to effectively learn the correspondence between observed flashlight images and the underlying geometry and material from synthetic datasets, enabling it to predict globally correct SDF and initial texture on a new scene, which eases the optimization process.

Formally, let K denote the number of input images, and let I_k represent the k-th input image. The model first extracts image features V_k from each input view k using a shared feature extractor $\mathcal{F}$.

$$V_k = \mathcal{F}(I_k), \tag{2}$$

Then, for a 3D point x, we project it onto each input view k by a perspective camera model π_k and get the image space location and depth by $(X_k, z_k) = \pi_k(x)$. The image feature vector is sampled by X_k in a pixel-aligned way and concatenated with the depth feature along the feature channel to form the complete feature vector f_k.

Given multi-view features at a point, there are multiple choices to fuse them, e.g. simple average [5], input mean and variance [48], and self-attention. Here we adopt the self-attention mechanism [14] due to its effectiveness in capturing geometry details from multi-view features.

Let $\mathcal{T}_{geo}$ and $\mathcal{T}_{app}$ be the geometry transformer and appearance transformer, respectively. We fuse the features from different sources with the two transformers:

$$f_{geo}, f_{app} = \mathcal{T}_{geo}(\{f_k\}), \mathcal{T}_{app}(\{f_k\}). \tag{3}$$

The fused feature vectors are then mapped to the signed distance d and the diffuse albedo ρ, the roughness γ, and specular albedo κ by two individual MLPs S_{θ_S} and P_{θ_P}:

$$d = S_{\theta_S}(f_{geo}), \tag{4}$$

$$\rho, \gamma, \kappa = P_{\theta_P}(f_{app}) \tag{5}$$

Loss functions. We define our loss functions in the pre-training stage as follows. The full loss L_{total} is composed of a geometry term L_{geo} and an appearance term L_{app}.

$$L_{total} = L_{geo} + L_{app}, \tag{6}$$

where the SDF is supervised by the geometry term L_{geo} which consists of four terms,

$$L_{geo} = L_{sdf} + \lambda_n L_{nml} + \lambda_e L_{eik} + \lambda_a L_{aln} \tag{7}$$

where $\lambda_n, \lambda_e, \lambda_a$ represent the weight of the corresponding loss terms. We adopt a binary cross-entropy loss and a surface constraint loss to supervise the signed distance field. Specifically, we sample N_l points near the surface and in the unit cube and use the label $l_i \in \{0, 1\}$ for points inside or outside the surface. We additionally sample N_s points on the surface and encourage the SDF at surface points to be zero. Finally, we compute L_{sdf} as:

$$L_{sdf} = \frac{1}{N_s}\sum_{i=1}^{N_s}|d_i| + \frac{\lambda_l}{N_l}\sum_{i=1}^{N_l} BCE(\phi(e^{\tau} d_i), l_i), \tag{8}$$

where ϕ denotes the sigmoid function, τ is a learnable parameter that controls the smoothness of the surface, and λ_l denotes the relative weight that balances the two terms.

We also supervise the SDF normal $\boldsymbol{n}$ by the ground truth normal $\hat{\boldsymbol{n}}$ on the sampled surface points.

$$L_{nml} = \frac{1}{N_s}\sum_{i=1}^{N_s}\|\boldsymbol{n}_i - \hat{\boldsymbol{n}}_i\|, \tag{9}$$

where $\boldsymbol{n}$ is defined by the normalized gradient of the estimated signed distance $\nabla d / \|\nabla d\|_2$.

To encourage the field to present a valid SDF, following IGR [49], we also adopt the geometric regularization loss L_{eik} on the sampled N_l space points.

$$L_{eik} = \frac{1}{N_l}\sum_{i=1}^{N_l}(1 - \|\nabla d_i\|_2)^2 \tag{10}$$

Additionally, we use a level set alignment loss L_{align} to encourage a better gradient consistency over the space [50].

$$L_{aln} = \frac{1}{N_l}\sum_{i=1}^{N_l}\beta_i(1 - \frac{\nabla d_i \cdot \nabla d_j}{\|\nabla d_i\|_2 \cdot \|\nabla d_j\|_2}), \tag{11}$$

where d_j denotes the signed distance of the point x_j, which is the projections of the query point x_i to the zero level set, β_i is an adaptive weight indicating the importance of x_i, which is given by $\beta_i = exp(-\delta|d_i|)$, where δ is set to 10 in our experiments.

To calculate the appearance loss L_{app}, we follow [5] by sampling N_c points around surfaces and penalize the L_1 distance between the predicted diffuse albedo and ground truth.

$$L_{app} = \frac{1}{N_c}\sum_{i=1}^{N_c}|\rho_i - \hat{\rho}_i|. \tag{12}$$

3.3 Joint Optimization

Although the pre-trained model can well capture globally correct geometry, directly applying the pre-trained model on in-the-wild data may inevitably result in local errors such as inaccurate geometry or materials, due to the non-neglectable difference between synthetic rendered images and real images in complex shading effects and the error-prone calibration process. The features learned from the

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591



Fig. 3. Qualitative comparison on the synthetic data. The geometry reconstruction results produced by our method exhibit much sharper geometry details and preserve thin structures while containing fewer artifacts such as missing body parts or severe distortions compared to the competitors. The diffuse albedo produced by ours contains less baked-in shadows. Please zoom in for details.

synthetic data may fail to represent the complex shading effects in real photos. Therefore, we propose to use differentiable physics-based rendering to jointly optimize the whole pipeline, using the supervision from the sparse-view flashlight images. We use sphere tracing algorithm [51] to efficiently locate the ray-surface intersections x_0. Note that x_0 is irrelevant to the network parameters. To propagate gradients to the network parameters during the optimization process, we use x instead of x_0 to compute the final colors. Here, x is an approximation of x_0 around surfaces and is made differentiable to network parameters Θ following IDR [21] by incorporating the surface normal $\boldsymbol{n}$:

$$x = x_0 - \frac{\boldsymbol{v}}{\boldsymbol{n} \cdot \boldsymbol{v}} d_0, \tag{13}$$

where x_0 denotes intersection points obtained from sphere tracing, $\boldsymbol{v}$ denotes the view direction, d_0 denotes the predicted signed distance at x_0.

Visual hull based culling. Despite the effectiveness of SDF as the implicit representation, a valid signed distance field is required to perform accurate sphere tracing. Therefore, it is crucial to properly initialize the neural network, such that the initial field approximates a signed distance field. A common practice in previous works [19], [21], [29] is to approximate the SDF of a unit sphere by properly initializing the MLP weights [52]. Then, they perform sphere tracing starting from the ray-sphere intersections. However, it is unsuitable in our case, as we observe that the neural fields directly predicted by the pre-trained network may not be a perfect SDF, particularly in the free space far away from the human body, due to a lack of supervision in the pre-training. In such cases, rays may pass through some thin structures or encounter undesired surfaces due to the wrong SDF value at the previous ray marching step, resulting in unstable optimization. In contrast, we observe that the SDF

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

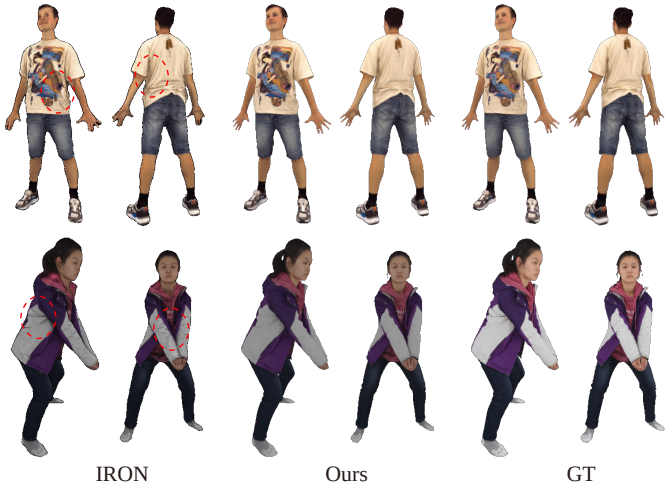


Fig. 4. Comparison of the reconstructed diffuse albedo. Our method recovers more accurate albedo with less baked-in shading.

predictions near the surface could be much more reliable. Based on the observations, given multi-view foreground masks M_k, we utilize the visual hull technique to mask out most of the empty regions to improve the speed and robustness of sphere tracing. Specifically, we first perform dilation operations on the masks and construct a dilated visual hull that safely contains the subject. Then, we precompute the front and back intersections of each ray with the dilated visual hull before optimization. In this way, we could perform sphere tracing from the front and back intersection of the visual hull, which both stabilizes the optimization process and reduces the computation cost.

Physics-based surface rendering. Given a surface location x determined by sphere tracing, we query the BRDF MLP to get the BRDF parameters and calculate shading by the rendering equation [53]. Mathematically, we compute the outgoing radiance L_o by:

$$L_o(\omega_o, x) = \int_\Omega L_i(\omega_i, x) f_r(\omega_i, \omega_o, x)(\omega_i \cdot n) d\omega_i, \quad (14)$$

where $L_i, \omega_o, \omega_i, n, f_r$ represent incident light, view direction, light direction, surface normal, and BRDF, respectively.

With the co-located flashlight and camera, the light direction ω_i is aligned with the view direction ω_o. We assume there is only one point light source and neglect the global illumination effect. Therefore, we could remove the costly integral on the hemisphere and use the simplified version of (14) as:

$$L_o(\omega_o, x) \approx L_i(\omega_i, x) f_r(\omega_i, \omega_i, x)(\omega_i \cdot n), \quad (15)$$

where the flashlight is modeled as a near field white point light source similar to the approach described in [29], [31]:

$$L_i(\omega_i, x) = \frac{L}{||x - o||_2^2}, \quad (16)$$

where L represents the light intensity which is modeled by a single learnable scalar tensor in each scene and o refers to the location of the light source.

As for the BRDF term f_r, we use the microfacet model [54] to represent the complex lighting effects in real data,

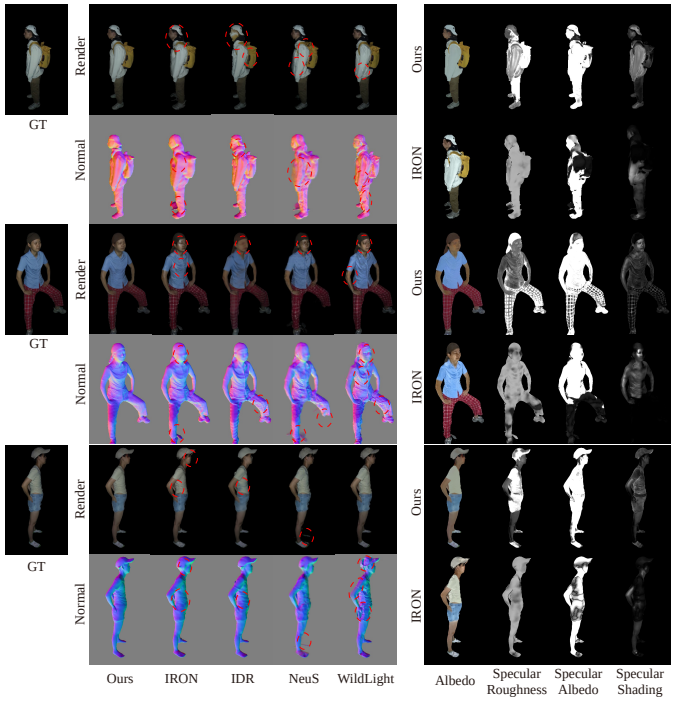


Fig. 5. Qualitative comparison on real captured data. The leftmost column shows the reference image, the middle columns show the reconstructed images and geometry, and the right 4 columns show the diffuse albedo, roughness, specular albedo, and specular shading, respectively. Our method yields more geometry details than the competitors while recovering diffuse albedo with less shading baked in and better specular maps. Zoom in for better details.

and a simple Lambertian BRDF is used on the synthetic data, due to the absence of specular maps in public available human scan dataset. The parameters required for the rendering model are encoded in different output channels of P_{θ_P}. Please refer to the supplementary materials for more details about the reflectance model.

The total loss at the optimization stage could be written as follows:

$$L_{total} = L_c + \lambda_m L_{mask} + \lambda_e L_{eik} + \lambda_a L_{aln} + \lambda_r L_{sp} + \lambda_t L_{smt}, \quad (17)$$

where $\lambda_m, \lambda_e, \lambda_a, \lambda_r, \lambda_t$ represent the weight of the corresponding loss terms, respectively. Let Q be the mini-batch pixels sampled at each iteration and $\hat{C}_q, O_q$ be the RGB and mask values corresponding to a pixel $q \in Q$. Let Q^{in} be the subset of Q where intersections x_q have been found and $O_q = 1$, and Q^{out} be the subset of Q where no intersections have been found or $O_q = 0$. The color loss is defined as:

$$L_c = \frac{1}{|Q|} \sum_{q \in Q^{in}} |C_q - \hat{C}_q| \quad (18)$$

, where C_q denotes the color computed by Eq. 15 at the intersection point x_q.

The mask loss is defined as:

$$L_{mask} = \frac{1}{e^\tau |Q|} \sum_{q \in Q^{out}} BCE(\phi(e^\tau d_q), O_q), \quad (19)$$

where τ is the learnable parameter inherited from the pre-training stage.

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

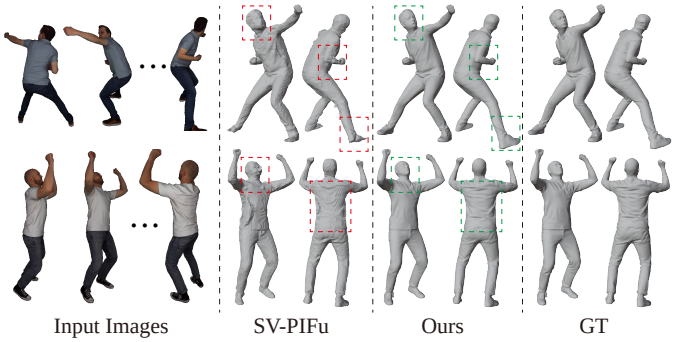


Fig. 6. Qualitative comparisons with SV-PIFu. Taking the same set of sparse view images as input, our approach reconstructs high-fidelity meshes with better geometry details and fewer artifacts.

The formulation of L_{eik} and L_{aln} are identical to the ones defined in the pre-training stage, except that they are calculated on the intersection points and uniformly sampled points inside the unit cube.

Due to the ill-posed setting of joint optimization of geometry and BRDF from sparse view, inspired by [45], we also adopt a sparsity loss L_{sp} and a smoothness loss L_{smt} to encourage the global sparsity and local smoothness of diffuse albedo.

The sparsity loss and smooth loss are defined as:

$$L_{sp} = gaussian_entrophy(\rho(x_q)) \tag{20}$$

$$L_{smt} = \frac{1}{|Q^{in}|} \sum_{q \in Q^{in}} |\rho(x_q) - \rho(x_q + \epsilon)| \tag{21}$$

where $gaussian_entrophy(\rho(x_q))$ is the entropy of the Gaussian distribution of the continuous diffuse albedo field as in [45], and ϵ is a 3D perturbation sampled from a Gaussian distribution $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 0, \sigma = 0.01$ in all experiments.

Once the optimization procedure converges, we can get the signed distance and appearance at any 3D location by querying the network. We use the marching cubes algorithm [36] to extract the final mesh and generate UV maps using Blender [55].

4 Experiments

4.1 Dataset

Synthetic Data. We use two public available high quality human scan dataset Thuman2.0 [56] as well as the CustomHuman Dataset [57] as benchmark to evaluate different methods. The Thuman2.0 dataset comprises 526 high-quality human scans of individuals, and the CustomHumans dataset has 643 scans, both encompassing a wide range of identities, outfits, and body poses. Our model is pre-trained using 290 human scan models from the Thuman2.0 dataset, following the train/test split introduced by [9], [58]. For each sample scan in the dataset, we use the Lambertian BRDF to render the images, due to the absence of specular maps in the dataset. Note that we treat the texture maps as the pseudo albedo maps for pre-training, as the true albedo maps are not available. We put a point light at the camera position as the only light source while rendering. We use a perspective camera model to render 60 images with associate alpha masks from 360° in the yaw axis with a step of 6°, each performing random augmentations on the view angle and the camera position to simulate the real capturing scenario. The resolution of the rendered images is 512×512.

TABLE 1
Quantitative comparison with SOTA methods on using 12 views on different datasets. Ours-fix denotes the variant of our method that keeps the image features fixed during optimization.

Data	Method	CD↓	NC↑	PSNR↑	SSIM↑	LPIPS↓
Thu [56]	IDR	3.556	0.926	28.786	0.974	0.0422
	Neus	5.711	0.894	28.976	0.965	0.0578
	IRON	4.859	0.903	31.703	0.978	0.0215
	Wild	7.354	0.864	27.257	0.978	0.0270
	SV-PIFu	4.954	0.895	-	-	-
	PIFu$^+$	7.728	0.909	27.382	0.959	0.0324
	Ours-fix	4.718	0.921	31.573	0.976	0.0216
	Ours	**1.944**	**0.950**	**35.373**	**0.989**	**0.0134**
Cus [57]	IDR	3.419	0.950	24.730	0.977	0.0393
	Neus	3.484	0.942	26.662	0.981	0.0356
	IRON	2.769	0.949	29.699	0.989	0.0140
	Wild	4.524	0.918	27.043	0.986	0.0182
	SV-PIFu	6.255	0.899	-	-	-
	PIFu$^+$	8.806	0.906	20.618	0.950	0.0428
	Ours-fix	4.180	0.936	27.500	0.981	0.0188
	Ours	**1.555**	**0.970**	**32.337**	**0.993**	**0.0096**
Real	IDR	-	-	28.848	0.898	0.1202
	Neus	-	-	29.292	0.879	0.1216
	IRON	-	-	29.728	0.894	0.0718
	Wild	-	-	29.606	0.914	0.0896
	PIFu$^+$	-	-	24.297	0.821	0.0986
	Ours-fix	-	-	29.364	0.883	0.0879
	Ours	-	-	**31.019**	**0.915**	**0.0503**

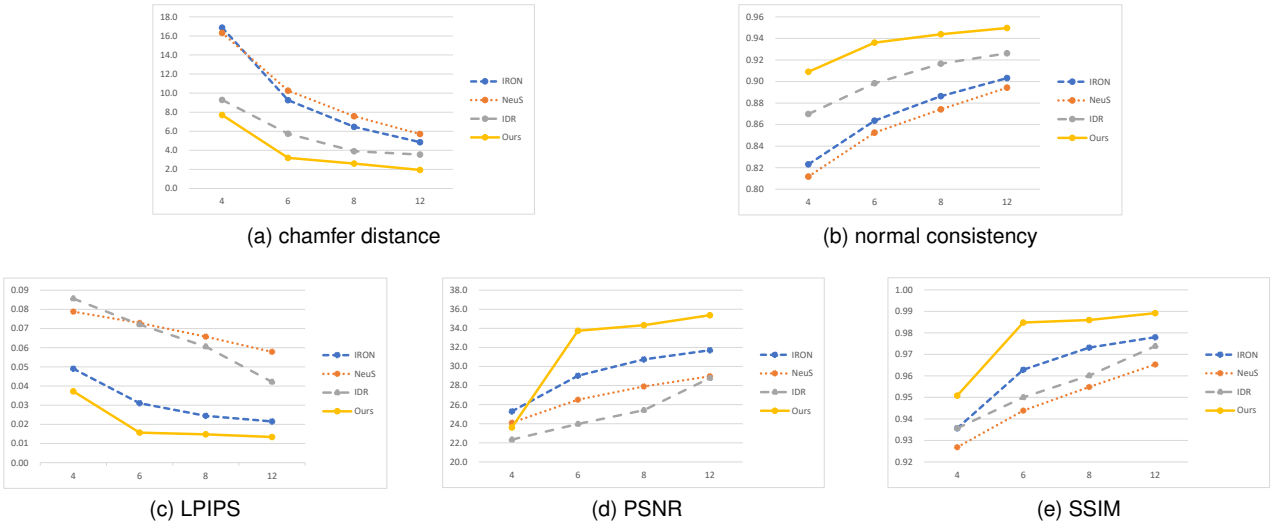
Real Data. We capture the real images of 6 volunteers in various poses and outfits in a dark environment using a camera equipped with an LED light source. Before capturing, we ask the participants to try to stay still without movements. Then, we move around the person to record a video with flash on and use Colmap [59], [60] to estimate the camera poses, which are further normalized so that the person is roughly located inside a unit sphere as in [29]. Finally, we uniformly sample frames from the video to form our target image set and use the off-the-shelf segmentation method [61] to obtain the foreground masks.

4.2 Training Details

Pre-training. To reduce the time consumption of loading and sampling mesh at training time, we pre-sample the points ahead of training. Specifically, we sample $200,000$ points on each mesh's surface and add random perturbations with a standard deviation $\sigma = 5cm$, and we sample another 50000 points uniformly in the unit cube to get a total of $250,000$ points in space. We set label 0 for points inside the surface and 1 for outside ones. Besides, we sample $200,000$ points on the surface and obtain the surface normal. At each training iteration, we randomly pick 4 views from the 60 rendered images of one subject as input, and we randomly draw $N_l = 5000$ and $N_s = 5000$ points from the pre-sampled points to predict the SDF and calculate the geometry loss L_{geo}. We utilize the Adam optimizer [62] and set the learning rate to be $1e-4$ and we set the loss weights as $\lambda_n = 0.1, \lambda_l = 0.2, \lambda_a = 0.01, \lambda_e = 0.1$, respectively. All

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591



(a) chamfer distance (b) normal consistency

(c) LPIPS (d) PSNR (e) SSIM

Fig. 7. Input image number evaluation. We experiment with different image numbers from 4 to 12 on the Thuman2.0 dataset. We find that our model significantly outperforms all the baseline methods for every metric across different numbers of views.



Fig. 8. Our reconstructed human mesh could be relit in various lighting conditions. The left two columns illustrate the relighting results with different point light locations. The right column shows the relighting results under different HDR environments. Our method produces realistic relighting results thanks to the detailed geometry and material.

network components are trained end-to-end for a day using a single NVIDIA RTX3090 GPU with 24G memory.

Joint optimization. With the pre-trained model, we joint optimize the model with sparse flash images on both synthetic and real data. We randomly sample 1024 rays from pixels inside the dilated mask from all the images. Due to the absence of specular textures in the existing human scan dataset, we optimize the surface albedo (the first three output channels of P_{Θ_P}) using the Lambertian BRDF model. For real data, we jointly optimize the diffuse and specular components using the microfacet model. In both cases, we employ the Adam optimizer with a learning rate of $1e-4$ which exponentially decays to $1e-5$ in 1000 iterations. The loss weights are set as $\lambda_m = 100, \lambda_e = 0.1, \lambda_a = 0.01, \lambda_r = 0.001, \lambda_t = 0.01$, respectively. We perform joint optimization for $\sim$1000 iterations which takes 10 minutes on average.

4.3 Metrics

We quantitatively evaluate the accuracy of our reconstruction using two geometry-related metrics (Chamfer distance(CD) and normal consistency(NC)) and three image-related metrics(PSNR, LPIPS, and SSIM). The Chamfer distance measures the global distance between point clouds, normal consistency is adopted to better reveal the quality of the geometry details. The three image-related metrics measures image quality and perceptual correctness of relighting and novel view synthesis tasks. The geometry metrics are reported on synthetic data only and the image metrics are reported on both synthetic and real data.

4.4 Results and Comparisons

Competing Methods. We choose four popular optimization-based methods NeuS [19], IDR [21], IRON [29], WildLight [30] and two learning based methods SV-PIFu [13] and PIFu$^+$ to compare with. SV-PIFu is a SOTA learning-based human reconstruction method that takes sparse views as input and fuses multi-view features with a transformer. As the original PIFu [5] can not reconstruct BRDF, we implement PIFu$^+$ by adding an MLP to predict the BRDF parameters. IDR is a pioneering work that reconstructs geometry and view-dependent appearance by a neural implicit field from multi-view images and the associate masks. NeuS is a popular framework after IDR that combines SDF and the volume rendering technique to recover the geometry and view-dependent appearance. IRON is an inverse rendering method that builds upon NeuS and IDR. It serves as a strong baseline that leverages the flashlight setting for geometry and BRDF reconstruction. We also compare our method with a recent work WildLight that performs shape and material reconstruction aided by flashlight. Note that their

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

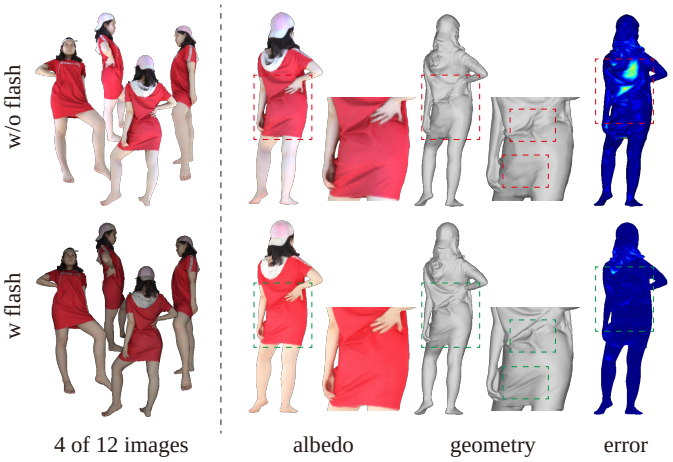


Fig. 9. Ablation of flashlight. The upper part shows the optimized geometry and color using images captured under constant lighting. The lower part shows the geometry optimized from images with the same camera views but captured using flashlight. The last column visualizes the geometry error by point-to-mesh distance. As can be viewed from the figure, the flashlight not only helps to recover the albedo but also enhances the geometry details.

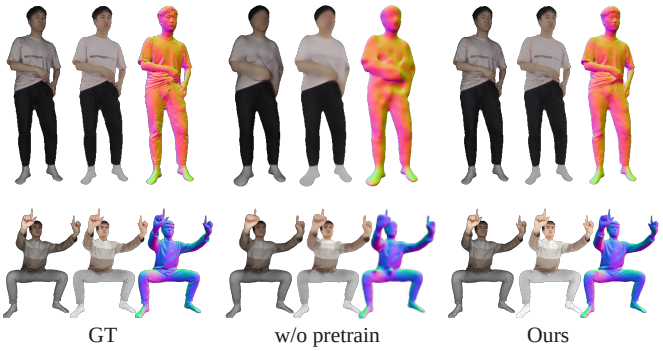


Fig. 10. Ablation of pre-training. Each group is represented by three columns, showing the rendered image, albedo, and normal. The variant without pre-training yields less detailed reconstructions.

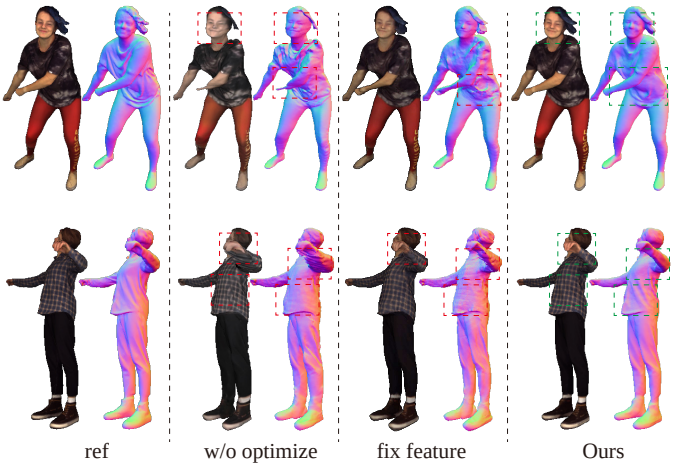


Fig. 11. Results directly obtained from the pre-trained model could contain artifacts, such as broken arms, twisted geometry, or inaccurate wrinkles, due to the train/test domain gap. The errors could be partially solved by fixing the image features and optimizing the transformers and MLPs, however, there are still artifacts on the local details. The joint optimization of the feature extractor successfully rectifies these artifacts.

capture setting is slightly different from ours as they also predict ambient lighting. For a fair comparison, we use the same images captured in a darkroom for WildLight. We retrain all the compared methods on our dataset and feed the foreground masks to all the competitors for a fair comparison.

Quantitative Results. We randomly select 20 models from the test set of the Thuman2.0 dataset and CustomHuman dataset and capture 6 real models to perform quantitative comparisons. The metrics are reported in Table 1. Our model produces the most accurate 3D geometry across all the competitors, as indicated by the smallest Chamfer Distance among all the methods. Furthermore, our approach obtains the best geometry details, i.e. surface normal, as can be proved by the highest normal consistency. Besides, our approach achieves the best score on the novel view synthesis task with light co-located at the camera position. This is mainly attributed to the better decoupling of shape and appearance by our pipeline. Note though our approach maintains a leading position among the competitors, we get a lower score of image metrics on the real captured data. We attribute this to the complexity of the real capturing scenario that fails to meet our assumptions, such as imperfect camera calibration, non-ideal point light source, and complex lighting effects such as subsurface scattering and global illumination.

Qualitative Results. We provide qualitative comparisons on the reconstruction quality of synthetic data in Fig. 3, Fig. 6, and real data in Fig. 5. Thanks to the physics-based optimization stage which addresses the local errors by joint optimizing the feature extractor, our method significantly outperforms the learning-based method PIFu$^+$ by the robustness of reconstruction, see the missing fingers and the broken arms of results by PIFu$^+$ in the first column of Fig. 3. Compared to SV-PIFu, meshes recovered by our approach exhibit fewer artifacts and more fine-grained details, e.g. human faces and garment wrinkles in Fig. 6. Meanwhile, Our method consistently produces sharper geometry details and more correct global structures compared to IDR, NeuS, and IRON. Note how the thin structures (e.g., fingers of the man standing in T-pose), garment wrinkles (e.g., the detailed wrinkles on the girl's coat), and facial geometry are successfully reconstructed in Fig. 3, while the competitors fail to reconstruct correctly. In addition, our method faithfully reconstructs high quality albedos, with less baked-in shading compared to IRON, as shown in Fig. 4 and Fig. 5. WildLight could recover sharp geometry but exhibits severe artifacts in local regions, such as the dent regions in the man's clothing and stretched faces of the girl, as seen in Fig. 3. The superior reconstruction quality is also verified by the comparison of real data shown in Fig. 5. Our model well captures the shading cues from multiple flashlight images and consistently outputs more detailed geometry and better materials than the others. Note how the details of the man's backpack, the wrinkles on the blue shirt of the woman, and the facial geometry are reconstructed with high fidelity. Additional relighting results of our method in diverse lighting conditions are shown in Fig. 8. Our method is able to produce realistic relighting results under various

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

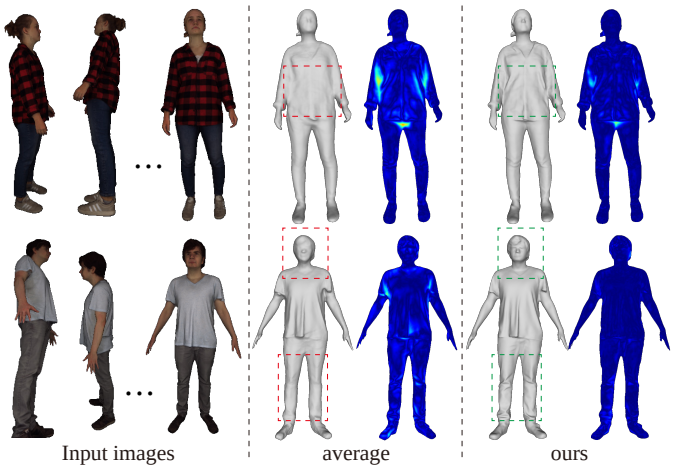


Fig. 12. Ablation of fusion methods, each group shows the reconstructed result and its geometric error. The baseline that fuses with averaging yields over-smoothed reconstructions compared with the transformer-based fusion.

TABLE 2
Quantitative comparison on the average reconstruction time of different methods on synthetic data.

Method	IDR	NeuS	IRON	WildLight	Ours
Time	∼2h	∼10h	∼8h	∼7h	**10min**

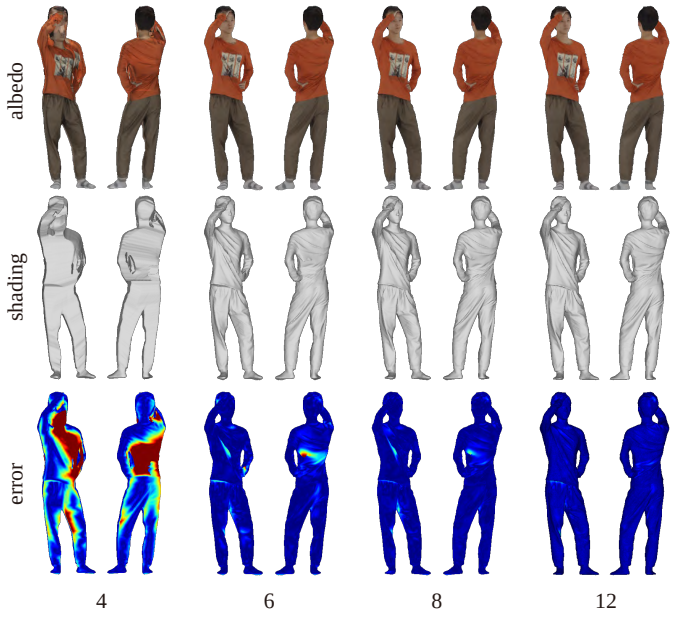


Fig. 13. We present different optimization results using 4,6,8, and 12 input views to demonstrate the impact of the number of views. With only 4 views, the optimization produces inferior results due to overfitting. When the input view number increases to 6, our model is able to generate relatively better results but with some local artifacts due to limited views in the local areas. Using 12 views as input achieves the highest accuracy in reconstructed geometry and albedo.

lighting conditions. Please refer to the companion video for more visualizations.

Time efficiency. We also analyze the reconstruction time of different optimization methods. We run each method using 12 views on 20 models randomly drawn from the Thuman2.0 and CustomHuman datasets and take the average time of reaching the same geometry accuracy. Thanks to the good starting point from pre-training, our method could produce detailed results in 10 minutes on average. IDR and NeuS both start optimization from an SDF that approximates the unit sphere by the initialization of MLPs [52]. Note that IDR is ∼4 times faster than NeuS, this is mainly due to the reduced computation cost of sphere tracing than the volume rendering with much fewer query points. IRON relies on a warm-up initialization from NeuS which takes an average of 4 hours, and the subsequent physics-based inverse rendering takes another 4 hours on average. WildLight takes around 7 hours as it performs volume rendering based on NeuS. All the experiments are performed on a single NVIDIA RTX 3090 GPU with 24G memory.

4.5 Ablations

Ablation of light. We evaluate the role of light in our pipeline by replacing the flashlights with constant environmental lighting. The experiment is conducted on the synthetic data to ensure the same camera poses and subjects except for lighting conditions. Instead of predicting BRDFs and calculating color by the physics-based renderer as in Eq. 14, we let the albedo predictor predict the color at point x and jointly optimize the geometry and color. The flashlight images provide view-dependent shading cues and facilitate geometry reasoning, therefore enhancing the accuracy, see the detailed wrinkles and the lower error in Fig. 9. In contrast, under static environment lighting, there is no way to decouple the shading and albedo, therefore, the model without flash images input produces light-entangled albedo and generates worse geometry.

Ablation of pre-training. We evaluate the importance of the pre-training stage by removing the learned image features and feeding the positional encoding of point location instead of depth. Instead of initializing SDF using a pre-trained model, we adopt the initialization method [52] in this experiment. Without pre-trained features on synthetic data, the optimization lacks a good start point, thereby getting degraded results and producing over-smoothed geometry and appearance, as shown in Fig. 10.

Ablation of image numbers. We evaluate the effect of the number of images on the performance of our method. We optimize our model using 4,6,8,12 images on synthetic data. With only 4 views, the optimization tends to overfit due to insufficient constraints, achieving an almost zero rendering loss from each view by predicting a surface field that forms a visual hull, with each surface point colored according to the RGB value of the corresponding pixel in the input image. As shown in Fig 14, the PSNR on training views continues to improve but declines on test views with further optimization. The overfitting issue diminishes as the number of images increases, as shown

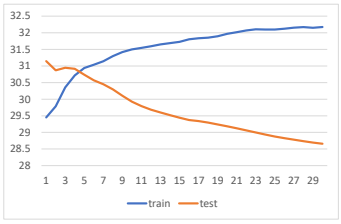


Fig. 14. PSNR on training and testing data when optimizing using only 4 views.The horizontal axis represents the number of iterations($\times 10$).

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

in Fig. 13, due to the additional constraints imposed on the field. The quantitative results can be viewed in Fig. 7, our method consistently outperforms the baselines on both geometry and appearance.

Ablation of optimization. We evaluate the effect of optimization in the pipeline by ablating the optimization scheme in two ways: 1) simply remove the optimization process and directly predict the fields (w/o optimize). 2) fix the feature extractor and optimize the other network components (ours-fix). The results can be found in Fig. 11 and Table. 1. As can be seen in the figure, the reconstruction result from a single forward pass of the pre-trained model without optimization contains local artifacts, such as the missing hand and distortions in the clothes in row 1, mainly due to the diversity in testing data that can not be fully covered by training data. If we fix the feature extractor and only optimize the MLPs and transformers, the results are still unsatisfactory, e.g. the distorted arms of the woman and the wrongly predicted geometry of the man, as shown in the third column of Fig. 11. With the proposed approach that jointly optimizes the feature extractor, the features could be quickly adapted to the new data, thus the issues are mitigated.

Ablation of fusion methods. We also evaluate the effect of the transformer-based fusion by comparing it with average-based fusion. A visual comparison of these two approaches can be viewed in Fig. 12. When features are averaged, a degradation in geometry details can be observed in the reconstruction results compared to transformer-based fusion. This is expected as the averaging operation does not consider the relative importance among views, leading to a loss of variance in features.

5 Limitations

The primary limitation of our study lies in the assumption of a dark environment with no other light sources. It would be challenging to combine flashlights in a normal environment in relightable human reconstruction, and we leave this as a future work. Another limitation lies in the assumption of a static scene during the capture, which restricts the person from moving during the capturing process. This constraint could be loosened if we manage to accelerate the capturing process, for example, one could use sparse pre-calibrated high-speed cameras and lights (e.g. DSLR rigs equipped with synchronized flashes) instead of capturing a video with a single camera, which sacrifices the usability for ordinary users, but still cheaper than the traditional light stages. In the future, we hope to address the movements by reconstructing a canonical space and adding a per-view residual motion field to fit the scene observation. Besides, our method relies on accurate foreground masks to perform optimization in the second stage. Failures in the camera pose estimation process could also lead to blurry artifacts in the reconstructed mesh. This issue could be alleviated by carefully placing objects with rich textures in the background to improve the accuracy of camera pose estimation.

6 Conclusion

This paper proposes a system for human detailed shape and intrinsic appearance reconstruction from images captured with a consumer camera with a co-located light. Our system achieves realistic relighting by effectively decoupling and recovering the body geometry and underlying material, with the help of shading cues from multi-view flashlight images. By integrating a pre-trained model in a differentiable physics-based rendering framework and performing a joint optimization, we achieve sparse-view reconstruction at a low computation cost on test data. The system has been verified on two synthetic human scan datasets and a real dataset captured in a darkroom. We believe our method provides a promising direction for high-quality human reconstruction with consumer-level devices. It's worth noting that there is a risk of misuse of our method, such as creating non-consensual 3D models or deepfakes. We firmly condemn any misuse of our work, especially when it results in the spread of false information or harms individuals' reputations.

7 Acknowledgement

The authors would like to thank the reviewers for their insightful comments. This work is supported by the National Key Research and Development Program of China under Grant 2022YFF0902302 and NSF China under Grants 62322209 and 62172357, and the 100 Talents Program of Zhejiang University.

References

[1] K. Ikeuchi, "Determining surface orientations of specular surfaces by using the photometric stereo method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 661–669, 1981.

[2] K. Guo, P. Lincoln, P. Davidson, J. Busch, X. Yu, M. Whalen, G. Harvey, S. Orts-Escolano, R. Pandey, J. Dourgarian *et al.*, "The relightables: Volumetric performance capture of humans with realistic relighting," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 6, pp. 1–19, 2019.

[3] T. Zhou, K. He, D. Wu, T. Xu, Q. Zhang, K. Shao, W. Chen, L. Xu, and J. Yu, "Relightable neural human assets from multi-view gradient illuminations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4315–4327.

[4] X. Zhang, S. Fanello, Y.-T. Tsai, T. Sun, T. Xue, R. Pandey, S. Orts-Escolano, P. Davidson, C. Rhemann, P. Debevec *et al.*, "Neural light transport for relighting and view synthesis," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 1, pp. 1–17, 2021.

[5] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2304–2314.

[6] S. Saito, T. Simon, J. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 84–93.

[7] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black, "Icon: Implicit clothed humans obtained from normals," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 13 286–13 296.

[8] R. Zheng, P. Li, H. Wang, and T. Yu, "Learning visibility field for detailed 3d human reconstruction and relighting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 216–226.

[9] K. Y. Chan, G. Lin, H. Zhao, and W. Lin, "Integratedpifu: Integrated pixel aligned implicit function for single-view human reconstruction," in *European conference on computer vision*. Springer, 2022, pp. 328–344.

[10] Z. Zheng, T. Yu, Y. Liu, and Q. Dai, "Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 3170–3184, 2021.

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

[11] T. Alldieck, M. Zanfir, and C. Sminchisescu, "Photorealistic monocular 3d reconstruction of humans wearing clothing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1506–1515.

[12] Z. Li, M. Oskarsson, and A. Heyden, "Detailed 3d human body reconstruction from multi-view images combining voxel super-resolution and learned implicit representation," *Applied Intelligence*, vol. 52, no. 6, pp. 6739–6759, 2022.

[13] P. Zins, Y. Xu, E. Boyer, S. Wuhrer, and T. Tung, "Data-driven 3d reconstruction of dressed humans from sparse views," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 494–504.

[14] Y. Zheng, R. Shao, Y. Zhang, T. Yu, Z. Zheng, Q. Dai, and Y. Liu, "Deepmulticap: Performance capture of multiple characters using sparse multiview cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6239–6249.

[15] R. Shao, H. Zhang, H. Zhang, M. Chen, Y.-P. Cao, T. Yu, and Y. Liu, "Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 872–15 882.

[16] https://www.daz3d.com/.

[17] https://renderpeople.com/.

[18] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[19] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 171–27 183, 2021.

[20] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9054–9063.

[21] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman, "Multiview neural surface reconstruction by disentangling geometry and appearance," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2492–2502, 2020.

[22] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3504–3515.

[23] M. Boss, V. Jampani, R. Braun, C. Liu, J. Barron, and H. Lensch, "Neural-pil: Neural pre-integrated lighting for reflectance decomposition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 691–10 704, 2021.

[24] M. Boss, R. Braun, V. Jampani, J. T. Barron, C. Liu, and H. Lensch, "Nerd: Neural reflectance decomposition from image collections," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 684–12 694.

[25] X. Zhang, P. P. Srinivasan, B. Deng, P. Debevec, W. T. Freeman, and J. T. Barron, "Nerfactor: Neural factorization of shape and reflectance under an unknown illumination," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 6, pp. 1–18, 2021.

[26] K. Zhang, F. Luan, Q. Wang, K. Bala, and N. 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*, 2021, pp. 5453–5462.

[27] P. P. Srinivasan, B. Deng, X. Zhang, M. Tancik, B. Mildenhall, and J. 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*, 2021, pp. 7495–7504.

[28] Z. Kuang, K. Olszewski, M. Chai, Z. Huang, P. Achlioptas, and S. Tulyakov, "Neroic: Neural rendering of objects from online image collections," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–12, 2022.

[29] K. Zhang, F. Luan, Z. Li, and N. Snavely, "Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5565–5574.

[30] Z. Cheng, J. Li, and H. Li, "Wildlight: In-the-wild inverse rendering with a flashlight," *arXiv preprint arXiv:2303.14190*, 2023.

[31] G. Nam, J. H. Lee, D. Gutierrez, and M. H. Kim, "Practical svbrdf acquisition of 3d objects with unstructured flash photography," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–12, 2018.

[32] F. Luan, S. Zhao, K. Bala, and Z. Dong, "Unified shape and svbrdf recovery using differentiable monte carlo rendering," in *Computer Graphics Forum*, vol. 40, no. 4. Wiley Online Library, 2021, pp. 101–113.

[33] https://polyhaven.com/.

[34] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8387–8397.

[35] F. Bogo, M. J. Black, M. Loper, and J. Romero, "Detailed full-body reconstructions of moving people from monocular rgb-d sequences," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2300–2308.

[36] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *Seminal graphics: pioneering efforts that shaped the field*, 1998, pp. 347–353.

[37] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, 2015.

[38] Y. Hong, J. Zhang, B. Jiang, Y. Guo, L. Liu, and H. Bao, "Stereopifu: Depth aware clothed human digitization via stereo vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 535–545.

[39] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black, "ECON: Explicit Clothed humans Optimized via Normal integration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[40] Y. Kanamori and Y. Endo, "Relighting humans: occlusion-aware inverse rendering for full-body human images," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–11, 2018.

[41] D. Tajima, Y. Kanamori, and Y. Endo, "Relighting humans in the wild: Monocular full-body human relighting with domain adaptation," in *Computer Graphics Forum*, vol. 40, no. 7. Wiley Online Library, 2021, pp. 205–216.

[42] M. Lagunas, X. Sun, J. Yang, R. Villegas, J. Zhang, Z. Shu, B. Masia, and D. Gutierrez, "Single-image full-body human relighting," *arXiv preprint arXiv:2107.07259*, 2021.

[43] R. Pandey, S. O. Escolano, C. Legendre, C. Haene, S. Bouaziz, C. Rhemann, P. Debevec, and S. Fanello, "Total relighting: learning to relight portraits for background replacement," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–21, 2021.

[44] C. Ji, T. Yu, K. Guo, J. Liu, and Y. Liu, "Geometry-aware single-image full-body human relighting," in *European Conference on Computer Vision*. Springer, 2022, pp. 388–405.

[45] Z. Chen and Z. Liu, "Relighting4d: Neural relightable human from videos," in *European Conference on Computer Vision*. Springer, 2022, pp. 606–623.

[46] U. Iqbal, A. Caliskan, K. Nagano, S. Khamis, P. Molchanov, and J. Kautz, "Rana: Relightable articulated neural avatars," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 142–23 153.

[47] Z. Xu, S. Peng, C. Geng, L. Mou, Z. Yan, J. Sun, H. Bao, and X. Zhou, "Relightable and animatable neural avatar from sparse-view video," *arXiv preprint arXiv:2308.07903*, 2023.

[48] T. Sun, K.-E. Lin, S. Bi, Z. Xu, and R. Ramamoorthi, "Nelf: Neural light-transport field for portrait view synthesis and relighting," *arXiv preprint arXiv:2107.12351*, 2021.

[49] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3789–3799.

[50] B. Ma, J. Zhou, Y.-S. Liu, and Z. Han, "Towards better gradient consistency for neural signed distance functions via level set alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 724–17 734.

[51] J. C. Hart, "Sphere tracing: A geometric method for the antialiased ray tracing of implicit surfaces," *The Visual Computer*, vol. 12, no. 10, pp. 527–545, 1996.

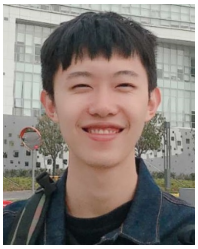
[52] M. Atzmon and Y. Lipman, "Sal: Sign agnostic learning of shapes from raw data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2565–2574.

[53] J. T. Kajiya, "The rendering equation," in *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, 1986, pp. 143–150.

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3450591

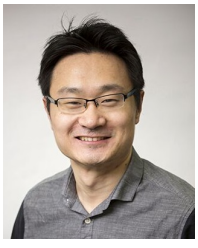
[54] B. Walter, S. R. Marschner, H. Li, and K. E. Torrance, "Microfacet models for refraction through rough surfaces," in *Proceedings of the 18th Eurographics conference on Rendering Techniques*, 2007, pp. 195–206.
[55] http://www.blender.org.
[56] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu, "Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5746–5756.
[57] H.-I. Ho, L. Xue, J. Song, and O. Hilliges, "Learning locally editable virtual humans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 024–21 035.
[58] K. Chan, G. Lin, H. Zhao, and W. Lin, "S-pifu: Integrating parametric human models with pifu for single-view clothed human reconstruction," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 373–17 385, 2022.
[59] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[60] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.
[61] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.
[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.



Jiawei Lu received his BS in Computer Science from Zhejiang University in 2017. He is working towards the PhD degree at the State Key Lab of CAD&CG, Zhejiang University. His research interests include human reconstruction, image-based modeling and deep learning.



Tianjia Shao received his BS from the Department of Automation, and his PhD in computer science from Institute for Advanced Study, both in Tsinghua University. He is currently a ZJU100 Young Professor in the State Key Laboratory of CAD&CG, Zhejiang University. Previously he was an Assistant Professor (Lecturer in UK) in the School of Computing, University of Leeds, UK. His current research focuses on 3D modeling from consumer hardware, and structure/function aware geometry processing.



He Wang received his PhD and did a post-doc in the School of Informatics, University of Edinburgh and his BS in Zhejiang University, China. He is an Associate Professor in the Virtual Environment and Computer Graphics (VECG) group, at the Department of Computer Science, University College London and a Visiting Professor at the University of Leeds. He is also a Turing Fellow and an Academic Advisor at the Commonwealth Scholarship Council. He serves as an Associate Editor of Computer Graphics Forum. His current research interest is mainly in computer graphics, vision and machine learning. Previously he was an Associate Professor and Lecturer at the University of Leeds UK and a Senior Research Associate at Disney Research Los Angeles.



Yong-Liang Yang received his BS and PhD of Computer Science from Tsinghua University. He is a Senior Lecturer in the Department of Computer Science, University of Bath. His research interests are broadly in visual computing and interactive techniques. He serves as Associate Editor of The Visual Computer and Journal of Economy and Technology. Previously he worked as a Post-doctoral Fellow and Research Scientist at King Abdullah University of Science and Technology.



Yin Yang received the PhD degree in computer science from the University of Texas at Dallas, in 2013. He is an associate professor with Kahlert School of Computing, University of Utah. He co-direct Utah Graphics Lab with his colleague Prof. Cem Yuksel. He is also affiliated with Utah Robotic Center. Before that, He was a faculty member at University of New Mexico and Clemson University. His research aims to develop efficient and customized computing methods for challenging problems in Graphics, Simulation, Deep Learning, Vision, Robotics, and many other applied areas.



Kun Zhou received the BS degree and PhD degree in computer science from Zhejiang University, in 1997 and 2002, respectively. He is a Cheung Kong professor with the Computer Science Department, Zhejiang University, and the director of the State Key Lab of CAD&CG. Prior to joining Zhejiang University in 2008, he was a leader researcher of the Internet Graphics Group, Microsoft Research Asia. He was named one of the world's top 35 young innovators by MIT Technology Review in 2011, was elected an IEEE Fellow in 2015 and an ACM Fellow in 2020. His research interests are in visual computing, parallel computing, human computer interaction, and virtual reality.

© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.