URHand: Universal Relightable Hands

Zhaoxi Chen[1,2*]   Gyeongsik Moon[1]   Kaiwen Guo[1]   Chen Cao[1]   Stanislav Pidhorskyi[1]

Tomas Simon[1]   Rohan Joshi[1]   Yuan Dong[1]   Yichen Xu[1]   Bernardo Pires[1]

He Wen[1]   Lucas Evans[1]   Bo Peng[1]   Julia Buffalini[1]   Autumn Trimble[1]

Kevyn McPhail[1]   Melissa Schoeller[1]   Shoou-I Yu[1]   Javier Romero[1]

Michael Zollhöfer[1]   Yaser Sheikh[1]   Ziwei Liu[2✉]   Shunsuke Saito[1✉]

[1]Codec Avatars Lab, Meta   [2]Nanyang Technological University

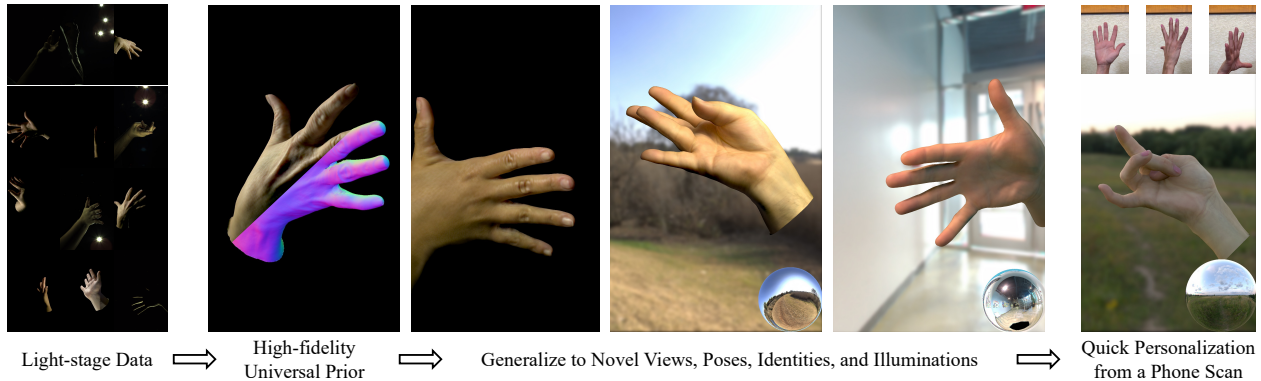https://frozenburning.github.io/projects/urhand

Figure 1. **URHand** (*a.k.a.* Your Hand). Our model is a high-fidelity **U**niversal prior for **R**elightable **Hand**s built upon light-stage data. It generalizes to novel viewpoints, poses, identities, and illuminations, which enables quick personalization from a phone scan.

Abstract

*Existing photorealistic relightable hand models require extensive identity-specific observations in different views, poses, and illuminations, and face challenges in generalizing to natural illuminations and novel identities. To bridge this gap, we present **URHand**, the first universal relightable hand model that generalizes across viewpoints, poses, illuminations, and identities. Our model allows few-shot personalization using images captured with a mobile phone, and is ready to be photorealistically rendered under novel illuminations. To simplify the personalization process while retaining photorealism, we build a powerful universal relightable prior based on neural relighting from multi-view images of hands captured in a light stage with hundreds of identities. The key challenge is scaling the cross-identity training while maintaining personalized fidelity and sharp details without compromising generalization under natural illuminations. To this end, we propose a spatially varying linear lighting model as the neural renderer that takes physics-inspired shading as input feature. By removing non-linear activations and bias, our specifically designed lighting model explicitly keeps the linearity of light transport. This enables single-stage training from light-stage data while generalizing to real-time rendering under arbitrary continuous illuminations across diverse identities. In addition, we introduce the joint learning of a physically based model and our neural relighting model, which further improves fidelity and generalization. Extensive experiments show that our approach achieves superior performance over existing methods in terms of both quality and generalizability. We also demonstrate quick personalization of URHand from a short phone scan of an unseen identity.*

1. Introduction

We engage our hands for various tasks throughout the day, and they consistently remain within our field of view. This constant visibility of our hands makes them one of the most frequently seen parts of our body, playing a central role in self-embodiment. To seamlessly reproduce this experience for games or social telepresence, an ideal hand representation in a digital medium is photorealistic, personalized, and importantly, relightable for coherent appearance in any environment. Our objective is to enable the quick creation of such a hand model for any individual given lightweight input such as a phone scan, without going through an expensive capture process in a production studio (Figure 1).

Approaches to build a photorealistic relightable hand

---

*This work was done during an internship at Meta
✉Corresponding authors

model can be broadly categorized into one of two philosophies. On the one hand, physically based rendering models [24, 53] can generalize to various illuminations through costly offline path-tracing, but typically lack photorealism under a real-time constraint. Additionally, accurately estimating material parameters remains a non-trivial challenge from unconstrained inputs and the quality is often bounded by the expressiveness of the physical models. On the other hand, neural relighting [2, 14] recently achieves impressive photorealism in real-time by directly inferring the outgoing radiance from illumination conditions. However, for generalization to natural illuminations to be possible these approaches require expensive data augmentation with teacher-student distillation, where the student model learns to match with offline renderings produced by the teacher model under natural illuminations. Most importantly, cross-identity generalization remains an open problem in both camps.

In this work, we propose URHand, the first **U**niversal **R**elightable **Hand** model that generalizes across viewpoints, motions, illuminations, and identities. To achieve the best trade-off between generalization and fidelity, our work exploits both physically based rendering and data-driven appearance modeling from neural relighting. More specifically, we incorporate known physics, such as the linearity of light transport [8] and surface reflections in the inductive bias of the neural relighting framework. We modulate non-linear layers conditioned by pose and identity with linear layers conditioned by spatially varying physically based shading [3]. This explicitly ensures linearity between input lighting features and output radiance. Thus, it enables environment map relighting without an expensive two-stage teacher-student distillation process commonly used in existing models [2, 14]. Our single-stage training enabled by linearity preservation makes cross-identity training more scalable with better generalization to novel illuminations.

Furthermore, we observe that the quality of the input shading features directly influences both generalization and fidelity of the final neural relighting outputs. Inspired by recent inverse rendering techniques [5, 60, 63], we introduce an additional physical branch that estimates the material parameters and high-resolution geometry via inverse rendering, from which we produce the input lighting features to the neural branch. The physical branch prevents the neural branch from overfitting by reducing hallucinations, and the neural branch compensates for the complex global light transport effects, such as subsurface scattering, that cannot be well captured by the physical branch. In addition, the proposed physics-based refinement improves the accuracy of the tracking geometry with fine details such as wrinkles. Combining it with our novel lighting-aware adversarial loss, our method achieves highly detailed relighting with various illuminations under any pose for novel identities.

We run an extensive ablation study as well as compar-

isons with baseline methods. The experiments demonstrate the efficacy of our hybrid neural-physical relighting method by outperforming other methods quantitatively and qualitatively. We also demonstrate the quick personalization of URHand from a phone scan, and relighting with arbitrary natural illuminations. In summary, our contributions are:

- The first method to learn a universal relightable hand model that generalizes to novel views, poses, illuminations, and identities.
- A spatially varying linear lighting model that generalizes to continuous illuminations without expensive distillation, enabling high-fidelity neural rendering and scalable training with multiple identities.
- A hybrid neural-physical relighting framework that leverages the best of both approaches to achieve high fidelity and generalization at the same time.
- The quick personalization of our universal prior to create a photorealistic and relightable hand from a phone scan.

2. Related Work

3D Hand Modeling. Human hand modeling is an active research field within vision and graphics. Early works focus more on 3D geometry and representation including mixture of 3D Gaussians [43, 44], sphere meshes [48], and triangular meshes [1, 7, 40]. These parametric hand models facilitate 3D hand pose estimation from 2D observations [27, 28, 31–33]. Recent works also incorporate physical priors [24, 29, 42, 49, 64] to model more accurate non-rigid deformation and articulation of the hand geometry. The recent advances of neural fields [47, 54] also enable the learning of personalized articulated models [18]. Beyond geometry modeling, achieving a lifelike appearance for hands [38] is paramount for realistic rendering and animation. Handy [37] learns a texture space by using generative adversarial networks for better photorealism and generalization. More recently, some methods [4, 6, 34] showcase the modeling of animatable hands from monocular/multiview captures. However, the appearance models of these approaches simply bake the captured illumination and cannot be rendered under novel illuminations. NIMBLE [24] builds PCA reflectance maps including diffuse, normal, and specular maps from light-stage data. DART [9] also supports accessories. However, physically based materials are expensive to render with global illumination which often limits their rendering fidelity with a real-time constraint. RelightableHands [14] enables the photorealistic relighting of hands in real-time using a neural appearance model. However, the method only supports person-specific modeling and generalization to unseen identity is not possible. In contrast, our approach generalizes across poses, illuminations, and identities, supporting relightable hand modeling of unseen identities from a phone scan.

Image-based Relighting. Given the linearity of light trans-

port, Debevec *et al*. [8] propose to render human faces under novel illuminations by linear combinations of sampled reflectance fields from one-light-at-a-time (OLAT) captures. A series of follow-up work [26, 45, 52, 55] enable dynamic relighting and capture through learning-based approaches. Meanwhile, another line of work aims at intrinsic decomposition, enabling physically based rendering with disentangled geometry and reflectance in the image space [11, 12, 17, 21, 35, 41]. Recent advances in neural rendering [15, 36, 58] learn to match with ground truth using non-linear neural networks given the lighting features from a simple physical shading model. Yet, image-based relighting suffers from 3D inconsistency and flickering due to the lack of an underlying 3D representation.

Model-based Relighting. To address the lack of 3D consistency in image-based relighting, one can leverage a shared 2D parameterization [56, 62] or template models [2, 5, 14, 19] for model-based relighting. Most model-based relighting approaches [5, 13, 16, 39, 50] rely on the intrinsic decomposition of geometry and reflectance followed by a physically based appearance model. While they achieve generalization to novel conditions, the fidelity is typically limited due to the lack of expressiveness in the underlying parametric BRDFs. On the other hand, methods with neural renderers support complex global illumination effects learned from captured data under point lights. However, generalization to continuous environments requires the expensive teacher-student training framework [2, 14], which is difficult to scale to cross-identity training. On the contrary, our spatially varying linear network achieves generalization to any type of illumination without additional training. Concurrently, Yang *et al*. [57] propose a linear lighting model for face relighting that eliminates the need of teacher-student distillation. While the motivation is similar, we observe that their holistic light representation does not generalize well for hands due to drastic visibility change by articulation (see Sec. 5.3 for analysis).

3. Preliminary

Data Acquisition. We use a multiview capture system consisting of 150 cameras and 350 LED lights to capture dynamic hands with time-multiplexed illuminations by interleaving fully lit (all lights on) and partially lit every other frame. Instead of OLAT, our partially lit frames use $L = 5$ grouped lights to increase brightness and reduce motion blur as in [2, 14, 22]. Images are captured in the resolution of 4096×2668 at 90 fps. We first reconstruct per-frame 3D meshes using [10] and detect 3D hand keypoints using [23] followed by triangulation from fully lit frames.

For partially lit frames, we leverage spherical linear interpolation over the pose parameter of adjacent fully lit frames to obtain the hand pose of partially lit frames. Our dataset contains 93 different identities in diverse hand mo-

tions with an average of 42000 frames for each identity.

Linearity of Light Transport. To render subjects in arbitrary illuminations, natural illumination is treated as a linear combination of distant point lights given the linearity of light transport [8]. Specifically, given the appearance value $\mathbf{C}^i$ based on the i-th point light, the final color $\mathbf{C}$ is computed as a linear combination of all light sources, *i.e.* $\mathbf{C} = \sum_{i=1}^{L} b_i \mathbf{C}^i$, where L denotes the number of lights, and b_i is the intensity of each light. Given a light transport function $f(b) = \mathbf{C}$, we define it as linear w.r.t. b such that:

$$f(b) = f(\sum_{i=1}^{L} b_i) = \sum_{i=1}^{L} f(b_i), \quad b = \sum_{i}^{L} b_i. \quad (1)$$

Hand Geometry Modeling. We use UHM [30] to represent 3D hands based on a mesh template with vertex offsets. The 3D hand can be driven by linear blend skinning (LBS) and represent identity- and pose-specific deformations.

Specifically, we design an autoencoder to obtain accurate hand tracking and geometry. The encoder learns to predict identity-dependent latent codes and 3D hand poses from the input fully lit frames. Given an articulated generic mesh template $\bar{\mathcal{M}} = \{\mathcal{V}, \mathcal{F}, \mathcal{U}, \theta\}$ with vertices $\mathcal{V} \in \mathbb{R}^{n_\mathcal{V} \times 3}$, faces $\mathcal{F} \in \mathbb{R}^{n_\mathcal{F} \times 3}$, texture coordinates $\mathcal{U} \in \mathbb{R}^{n_\mathcal{V} \times 2}$, and 3D pose $\theta \in \mathbb{R}^{60 \times 3}$ in Euler angles and latent codes z, the decoder learns to predict the 3D offset of all vertices as $\delta\mathcal{V} \in \mathbb{R}^{n_\mathcal{V} \times 3}$. We use $n_\mathcal{V} = 15930$ and $n_\mathcal{F} = 32340$. The tracked hand mesh will be represented as $\mathcal{M} = \{\mathcal{V} + \delta\mathcal{V}, \mathcal{F}, \mathcal{U}, \theta\}$. Please refer to the supplementary material for more details.

4. Universal Relightable Hands

Our goal is to build a universal relightable shape and appearance model for human hands that can be rendered for any identity under arbitrary illumination in real-time. To this end, we learn a relightable appearance model from cross-identity light-stage captures based on grouped point lights.

In this section, we will introduce our learning framework, URHand, which learns to relight target hands in different poses and views. The core of our model is a spatially varying linear lighting model that preserves the linearity of light transport, which enables the generalization to arbitrary illuminations by training with monochrome group lights only. Our model consists of two parallel rendering branches, physical and neural. The physical branch (Sec. 4.1) focuses on refining geometry and providing accurate shading features as an illumination proxy for the neural branch. The neural branch (Sec. 4.2) learns the final appearance of hands with global illumination. These two branches are trained jointly in an end-to-end manner with our tailored loss functions (Sec. 4.3). Finally, we use this universal prior to quickly personalize a relightable hand model from few-shot observations (Sec. 5.4).
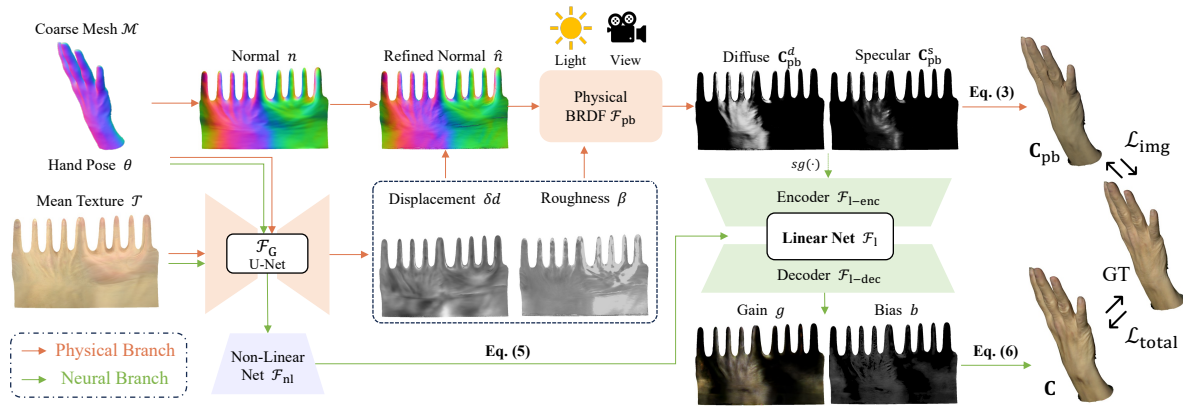
Figure 2. **Overview of URHand.** Our model takes as input a mean texture $\mathcal{T}$, hand pose θ, and a coarse mesh $\mathcal{M}$ for each identity. The physical branch (Sec. 4.1) focuses on geometry refinement and providing accurate shading features for the neural branch (Sec. 4.2). The core of the neural branch is the linear lighting model which takes as input the physics-inspired shading features from the physical branch. The neural branch learns to predict the gain and bias map over the mean texture. We leverage a differentiable rasterizer for rendering and minimize the loss of both branches against ground truth images (Sec. 4.3). The $sg(\cdot)$ denotes the stop-gradient operation.

4.1. Physically based Geometry Refinement

The physical branch employs online physically based rendering to render images using a parametric BRDF. We optimize the material parameters of the BRDF via inverse rendering. The goal of the physical branch is two-fold: 1) further refine the initial hand geometry for better alignment, and 2) provide generalizable lighting features that best approximate the specular reflection and diffuse shading to prevent overfitting in the neural relighting.

We illustrate the physical branch in Fig. 2. The physical branch estimates a parametric Disney BRDF [3] $\mathcal{F}_{\mathrm{pb}}$. We use a 2D U-Net $\mathcal{F}_{\mathrm{G}}$ to infer a displacement map $\delta d \in \mathbb{R}^{1024 \times 1024 \times 3}$ and a roughness map $\beta \in \mathbb{R}^{1024 \times 1024}$ in UV space. Instead of predicting a normal map directly, we use the predicted displacement map to update the normal map on top of the unwrapped coarse mesh in the UV space, which makes it easier to infer high-frequency geometric details on a smooth base surface [59]. To support cross-identity modeling with pose-dependent geometry and appearance change, $\mathcal{F}_{\mathrm{G}}$ takes as input the hand pose θ and the unwrapped mean texture $\mathcal{T} \in \mathbb{R}^{1024 \times 1024 \times 3}$ for each identity, where the hand pose is concatenated to the bottleneck representation of the U-Net.

Specifically, the refined surface $\hat{\mathbf{x}}$ is obtained by adding the displacement δd to the base surface $\mathbf{x}$ derived from the coarse mesh $\mathcal{M}$ along the direction of the normal $\mathbf{n}$:

$$\hat{\mathbf{x}} = \mathbf{x} + \delta d \cdot \mathbf{n}, \quad (2)$$

where positional map $\hat{\mathbf{x}}$ is then used to obtain the refined normal $\hat{\mathbf{n}}$ in the UV space for physically based rendering. In our implementation, we apply a sigmoid activation followed by a scaling factor of 3 in order to constrain the range of displacement to ± 3mm.

The refined normal $\hat{\mathbf{n}}$ and roughness β are fed into $\mathcal{F}_{\mathrm{pb}}$ that considers only the first bounce. Given an illumination $\mathcal{L}$, the final color $\mathbf{C}_{\mathrm{pb}}$ from camera view $\mathbf{d}$ is computed as:

$$\mathbf{C}_{\mathrm{pb}}(\hat{\mathbf{x}}, \mathbf{d}, \mathcal{L}) = \int L_i(\mathcal{L}, \hat{\mathbf{x}}, \omega_i) \mathcal{F}_{\mathrm{pb}}(\hat{\mathbf{x}}, \omega_i, \mathbf{d}, \beta)(\omega_i \cdot \hat{\mathbf{n}}) d\omega_i, \quad (3)$$

where $L_i(\hat{\mathbf{x}}, \omega_i)$ is the incident light from direction ω_i. The physically rendered texture map $\mathbf{C}_{\mathrm{pb}}$ can be decomposed into physically based shading feature, *i.e.* $\mathbf{F}_{\mathrm{pb}}(\mathcal{L}) = \{\mathbf{C}_{\mathrm{pb}}^d, \mathbf{C}_{\mathrm{pb}}^s\}$, according to the equation $\mathbf{C}_{\mathrm{pb}} = \mathbf{C}_{\mathrm{pb}}^d \odot \mathcal{T} + \mathbf{C}_{\mathrm{pb}}^s$, where $\odot$ is the element-wise multiplication, and $\mathcal{T}$ is the mean texture approximating albedo. Note that, this computation is directly performed in the UV space.

4.2. Linear Lighting Model

The rendering based on a parametric BRDF generalizes to novel illuminations, but they lack correct global light transport effects such as subsurface scattering. To enable relighting with global illumination in real-time, we introduce a neural renderer $\mathcal{F}_{\mathrm{R}}$. Given a target illumination, the outgoing radiance is computed as: $\mathbf{C} = \mathcal{F}_{\mathrm{R}}(\mathcal{M}, \mathcal{T}, \mathbf{F}_{\mathrm{pb}}(\mathcal{L}), \theta)$.

Our key insight is that removing the non-linear activation layers and bias in a convolutional neural network preserves the linearity with respect to the input features. Since our input feature is the physically based shading $\mathbf{F}_{\mathrm{pb}}(\mathcal{L})$, our network satisfies the following:

$$\sum_{i=1}^{L} \mathcal{F}_{\mathrm{R}}(\mathcal{M}, \mathcal{T}, \mathbf{F}_{\mathrm{pb}}(\mathcal{L}_i), \theta) = \mathcal{F}_{\mathrm{R}}(\mathcal{M}, \mathcal{T}, \sum_{i=1}^{L} \mathbf{F}_{\mathrm{pb}}(\mathcal{L}_i), \theta). \quad (4)$$

Since the physically based rendering in Eq. 3 is energy preserving, the network holds the linearity w.r.t. the input illumination. Therefore, our network can produce accurate relighting with continuous environment maps by training only with discrete point lights without additional distillation.

We illustrate the architecture of $\mathcal{F}_{\mathrm{R}} = \{\mathcal{F}_{\mathrm{l}}, \mathcal{F}_{\mathrm{nl}}\}$ in Fig. 2. It consists of a linear ($\mathcal{F}_{\mathrm{l}}$) and a non-linear ($\mathcal{F}_{\mathrm{nl}}$) branch, where the linear branch consists of an encoder $\mathcal{F}_{\mathrm{l-enc}}$ and a decoder $\mathcal{F}_{\mathrm{l-dec}}$ with physically based shading features $\mathbf{F}_{\mathrm{pb}}(\mathcal{L})$ as input. The pose- and identity-dependent features derived from the mean texture $\mathcal{T}$ and pose θ are fed

into a non-linear branch. We fuse the linear and non-linear feature maps in the decoder of the linear branch as follows:

$$\mathcal{F}_{l-dec}^{j+1} = \frac{1}{\sqrt{2}} \cdot \text{ConvT}(\mathcal{F}_{l-enc}^j + \mathcal{F}_{l-dec}^j) \odot \mathcal{F}_{nl}^j, \quad (5)$$

where j is the index of the layer, ConvT is the transposed convolutional layer without bias. This fusion mechanism keeps the linearity of the output w.r.t. the input lighting features while incorporating non-linearity w.r.t. identity and pose conditions. Instead of predicting the final texture, our neural renderer predicts the texel-aligned gain map g and bias map b, which contribute to the final texture as follows:

$$\mathbf{C} = g \odot \mathcal{T} + b \cdot \sigma_{\mathcal{T}}, \quad (6)$$

where $\sigma_{\mathcal{T}} = 64$ is the standard deviation of textures.

Important to note that since our input feature $\mathbf{F}_{pb}$ is spatially varying in a texel-aligned manner similar to [14, 36], we can accurately incorporate shadow information for better generalization with diverse poses. While the concurrent work [57] also proposes a linear lighting model using a holistic illumination representation by simply reshaping environment maps, we observe that the holistic illumination representation cannot generalize for hands due to the infinite shadow variations caused by articulation. The same observation is also reported in person-specific relightable hand modeling [14]. Please refer to Sec. 5 for the analysis.

4.3. Training Objectives

Our model is trained on multiview partially lit images in different identities and poses. The training objective $\mathcal{L}_{total}$ consists of three parts: reconstruction loss $\mathcal{L}_{img}$, lighting-aware adversarial loss $\mathcal{L}_{GAN}$, and L1 regularization $\mathcal{L}_{reg}$:

$$\mathcal{L}_{total} = \lambda_{img}\mathcal{L}_{img} + \lambda_{GAN}\mathcal{L}_{GAN} + \lambda_{reg}\mathcal{L}_{reg}, \quad (7)$$

where λ_* are corresponding loss weights.
Reconstruction Loss. We leverage a sum of L1 loss $\mathcal{L}_{MAE}$ and perceptual loss $\mathcal{L}_{eff}$ based on the EfficientNet [46] backbone, i.e. $\mathcal{L}_{img} = \mathcal{L}_{MAE} + \mathcal{L}_{eff}$. Both the renderings from the neural and the physical branch are supervised by the reconstruction loss against the ground truth images.
Lighting-aware Adversarial Loss. To improve the visual quality, we propose to use an adversarial loss on top of the reconstruction loss. We found that a naive image-conditioned discriminator performs poorly due to the significant appearance change in partially lit frames. To address this, we leverage a lighting-aware discriminator on multiple scales of renderings. Specifically, the discriminator $\mathcal{F}_D$ is conditioned on the diffuse and specular feature, $\{A, S\}$, which prompts the network to discriminate real and fake images given the illumination information. These lighting features are based on simple Phong reflectance [14] to ensure a consistent lighting prompt during training. We choose a

hinge loss [25] operated on multi-resolution discriminated patches as the adversarial target:

$$\mathcal{L}_{GAN} = \log \mathcal{F}_D(I|A, S) + \log[1 - \mathcal{F}_D(\hat{I}|A, S)], \quad (8)$$

where I is the ground truth, and $\hat{I}$ is the rendered image.
L1 Regularization on Linear Model. We also discovered that without regularization, a linear lighting model often produces noticeable flickering. As a linear convolutional network is capacity-limited, it tends to have high variance in intermediate features, resulting in poor generalization to novel illuminations. Thus, we penalize the L1 norm of the intermediate features from all layers in the linear branch:

$$\mathcal{L}_{reg} = \sum_{j=1}^{N} ||F_{l-enc}^j||_1. \quad (9)$$

Implementation Detail. The optimizable modules are $\{\mathcal{F}_G, \mathcal{F}_R, \mathcal{F}_D\}$. We use the Adam [20] optimizer and set the loss weights as $\lambda_{img} = 1.0$, $\lambda_{GAN} = 0.01$, and $\lambda_{reg} = 0.01$, respectively. We train the model for 2M iterations distributed on 8 A100 GPUs with a batch size of 24 in total. The initial learning rate is 1×10^{-4} which decays to 3×10^{-5} with a multistep learning rate scheduler. We describe the detailed network architecture and more training details in the supplementary.
Runtime Analysis. Our proposed linear lighting model is not only scalable and generalizable but also efficient for real-time rendering. Our model achieves 38 FPS (25.7 ms) given grouped lights as input and 31 FPS (31.9 ms) given environmental maps as input on an NVIDIA A100 GPU.

5. Experiments

5.1. Evaluation Protocols

To quantitatively evaluate the fidelity of each method, we use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure [51] (SSIM), and Learned Perceptual Image Patch Similarity [61] (LPIPS) as metrics. To solely evaluate the quality of rendered hands, we only take the foreground according to the mask obtained from the refined hand geometry. We exclude several segments from training data to evaluate the generalization of our model to novel poses. All metrics are evaluated on 1,000 images randomly sampled from those held-out segments.

For fair comparisons with existing works, we train and evaluate all methods in the single-identity setting except mentioned. We carefully held out several light directions and a motion segment from the training data for evaluation, and made sure training and test identities are in different genders and skin tones. For ablation studies, we train with ten identities and test on 1) unseen segments from a train subject, 2) unseen segments from an unseen subject, and 3) an unseen illumination from a train subject, which orthogonally evaluate the generalization of our model to novel poses, identities, and illuminations.

Table 1. **Quantitative comparisons on sequences with grouped lights**. We evaluate our method for both per-subject optimization and novel identity generalization against the state-of-the-art methods in model-based hand relighting. †Methods are evaluated on the training identity with unseen segments. *Methods are evaluated on unseen identity during training. The top three techniques are highlighted in red, orange, and yellow, respectively.

Method	Subject 1			Subject 2			Subject 3		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
†RelightableHands [14]	25.97	0.9301	0.1425	25.92	0.9372	0.1426	27.16	0.9419	0.1280
†Ours (Physical only)	23.44	0.9062	0.1708	23.25	0.9154	0.1715	24.90	0.9216	0.1510
†Ours (Full model)	27.77	0.9400	0.1204	26.36	0.9384	0.1344	27.75	0.9445	0.1226
*Handy [37] + Phong	16.65	0.8235	0.3026	16.39	0.8328	0.3019	17.70	0.8311	0.2872
*Handy [37] + GGX	16.60	0.8212	0.3125	16.33	0.8300	0.3107	17.58	0.8288	0.2978
*Ours (Full model)	26.94	0.9271	0.1335	26.24	0.9368	0.1341	27.58	0.9436	0.1197

a) Single Identity Evaluation   b) Novel Identity Generalization

RelightableHands   **Ours** (Physical Only)   **Ours**   Handy + Phong   Handy + GGX   **Ours**   Ground Truth
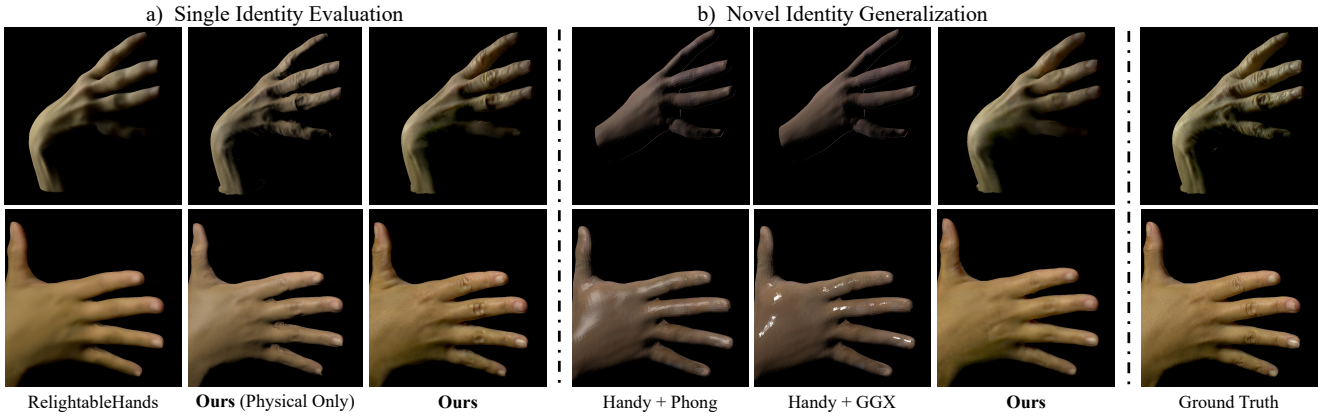
Figure 3. **Qualitative comparisons on sequences with grouped lights.** We evaluate our method for both per-subject optimization and novel identity generalization against comparison methods. a) All methods are evaluated on the training identity with unseen segments. b) Methods are evaluated on unseen identity during training.

5.2. Comparisons

We compare our approach with the state-of-the-art 3D hand relighting and reconstruction methods. Relightable-Hands [14] reconstructs relightable appearance via per-identity optimization. Handy [37] predicts the UV texture of the hand from in-the-wild images. For fair comparisons, we apply physically based renderers on top of the predicted texture to enable relightable appearances. We also compare with our physical branch using the same loss functions. This provides a fair comparison with physically based relighting. For this reason, we omit the comparison with other existing physically based relightable hand methods such as HARP [19] and NIMBLE [24]. Please refer to the supplementary for our detailed implementations of these methods.

We present quantitative results in Table 1. For per-identity training, our method significantly outperforms baseline methods on all metrics, which highlights the effectiveness of our key designs. As shown in Figure 3, our method is able to reproduce detailed geometry, *e.g.* wrinkles and nails, together with high-fidelity details like specularities and shadows. Handy [37] fails to reproduce the base texture of the target hand due to their data-driven latent tex-

ture space. Its combinations of simple physical renderers lead to artifacts of the relit results. RelightableHands [14] can reproduce correct shading and appearance. However, the quality of geometric details and specularity is not on par with the proposed method, which demonstrates the effectiveness of our hybrid neural-physical approach. Furthermore, the generalizability of our method is showcased in the right side of Figure 3, where the shown test subjects are withheld from the training set. Although the quality slightly degrades compared with per-identity optimization, it still outperforms all other baselines by a large margin.

5.3. Ablation Studies

Ablation on Linear Lighting Model. We investigate different designs of the linear lighting model. We compare our spatially varying linear lighting model with three alternatives: **1)** Non-linear model, where all non-linear activations are turned on; **2)** Non-linear model with a linearity consistency loss to constrain the linearity (Eq. 1) between output and input of the network; **3)** MLP-based linear model proposed in [57], where the lighting is represented as environment map and flattened to a 1-D vector.

Table 4 shows that the proposed spatially varying lin-

Table 2. **Ablation studies of the linear lighting model on sequences with grouped lights**. The "Non-linear" denotes the model that does not satisfy the linearity of light transport while the "Linearity Consistency" denotes the non-linear model with regularizations to constrain the linearity between output and input. Moreover, in contrast to MLP-based linear model [57], our model is a spatially varying linear model. The top three techniques are highlighted in red, orange, and yellow, respectively.

| Method	Trained Subject			Unseen Subject			Fully lit		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Non-linear	25.79	0.9283	0.1364	23.89	0.9145	0.1523	17.50	0.9070	0.1717
Linearity Consistency	24.77	0.9122	0.1560	23.11	0.9124	0.1537	19.75	0.9013	0.1619
MLP-based Linear [57]	22.19	0.8727	0.1885	21.56	0.8787	0.1823	11.56	0.8644	0.2259
Ours (Linear Model)	26.01	0.9270	0.1336	24.70	0.9121	0.1520	20.61	0.9153	0.1576
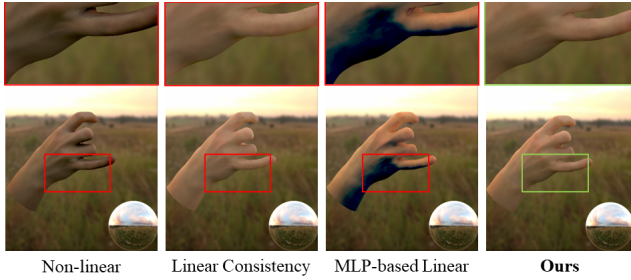
Figure 4. **Ablation study on the design of linear lighting model.** Our spatially varying linear lighting model produces realistic renderings, while the baseline methods fail to correctly model shadows or tend to be over smooth.

Table 3. **Ablation studies of the lighting features on sequences with grouped lights**. The top three techniques are highlighted in red, orange, and yellow, respectively.

Method	Trained Subject			Unseen Subject		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Phong	25.53	0.9257	0.1382	23.87	0.9148	0.1523
w/o Specular	24.97	0.9160	0.1498	23.20	0.9075	0.1607
w/o Visibility	25.44	0.9228	0.1426	23.75	0.9133	0.1533
w/o Refiner	24.82	0.9144	0.1437	22.80	0.8993	0.1620
Full Model	26.01	0.9270	0.1336	24.70	0.9121	0.1520

ear lighting model achieves the best performance on both trained and unseen identities. Moreover, we also quantitatively evaluate on fully lit frames to validate the generalization to novel illumination. The non-linear baseline fails to generalize, which shows the importance of linearity in our network. The MLP-based linear model fails to correctly model the light transport due to the lack of pose-aware visibility change. Figure 4 shows that the proposed lighting model produces the most realistic renderings, while the baseline methods struggle with correctly modeling shadows and detailed shading effects.

Ablation on Different Lighting Features. We also evaluate the effectiveness of our lighting feature representation, *i.e.* the diffuse $\mathbf{C}_{\mathrm{pb}}^d$ and specular feature $\mathbf{C}_{\mathrm{pb}}^s$. The contribution in the lighting features can be factored into **1)** the underlying BRDF, **2)** the visibility, and **3)** the geometry. As shown in Figure 5, our model reproduces specular highlights due to the more accurate specular feature from the optimized Disney BRDF. This indicates that the quality
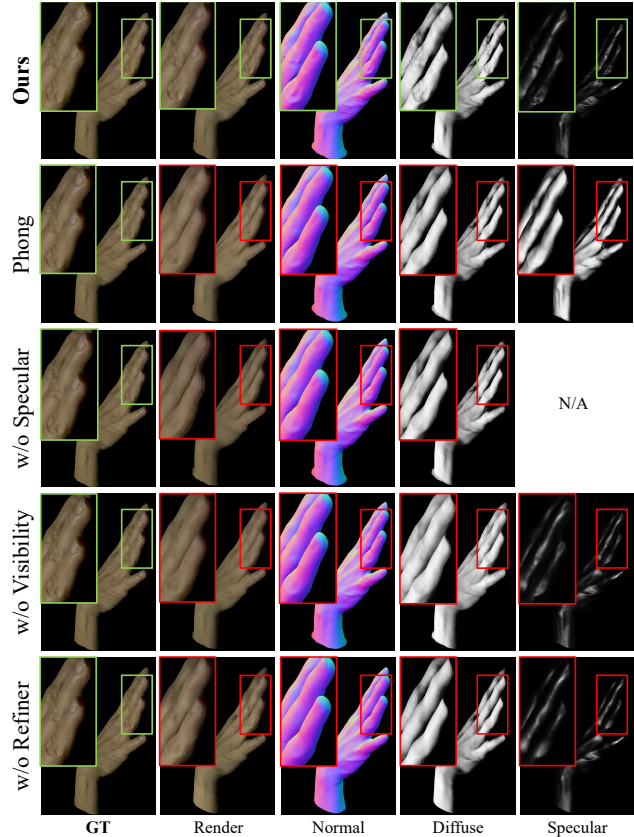

Figure 5. **Ablation studies on the impact of lighting features and geometry refinement.** Notably, our full model can produce fine-grained geometry like wrinkles and nails as well as specular highlights (*e.g.* little finger).

of the lighting feature determines the quality of the neural relighting. In fact, the neural relighting based on Phong specular features [14, 36] fails to produce accurate specular highlights. The visibility is also important for novel pose generalization. Compared with the model trained without visibility, our full model reproduces better soft shadows as well as detailed geometry. These observations are also supported by our quantitative evaluation presented in Table 3.

Effectiveness of Geometry Refinement. Our proposed neural-physical rendering offers the ability to refine geom-

Table 4. **Ablation studies of the proposed training objectives on sequences with grouped lights**. The top three techniques are highlighted in red, orange, and yellow, respectively.

Method	Trained Subject			Unseen Subject		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
w/o $\mathcal{L}_{GAN}$	25.95	0.9271	0.1392	23.70	0.9113	0.1542
w/o Light-aware $\mathcal{L}_{GAN}$	25.38	0.9215	0.1391	23.88	0.9142	0.1521
w/o L1 Reg $\mathcal{L}_{reg}$	25.52	0.9242	0.1351	23.90	0.9149	0.1491
Full Model	26.01	0.9270	0.1336	24.70	0.9121	0.1520
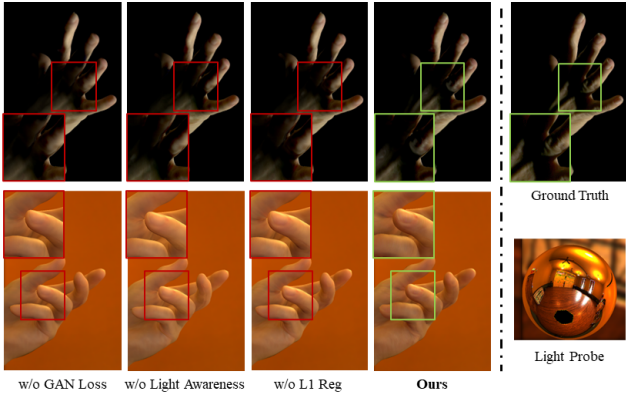
Figure 6. **Ablation studies on the proposed training objectives.** The adversarial loss improves the overall quality while the light awareness of the discriminator is critical for correct shadows.

etry. As shown in Figure 5, our full model can produce more fine-grained details such as wrinkles and nails. The geometry refinement evidently prevents the neural renderer from hallucinating, leading to better generalization to novel identities and illuminations. The quantitative results in Table 3 show that our geometry refinement (against w/o Refiner) significantly improves the reconstruction accuracy for both training and unseen identities. This validates the generalizability of our geometry refinement to novel identities.
Effectiveness of Adversarial Loss. We evaluate the effectiveness of the proposed lighting-aware adversarial loss. Quantitative results in Table 4 suggest that this adversarial loss improves the overall quality. The lighting-aware discriminator further improves the fidelity. Figure 6 validates that the adversarial loss is critical for reproducing shadow and detailed geometry. It also enhances specular highlights.
Effectiveness of L1 Regularization. We validate the effectiveness of the L1 regularization on the intermediate features of the linear lighting model. Table 4 and Figure 6 illustrate that, compared with the model without L1 regularization, our model achieves better modeling of the soft shadows and appearance. Its effectiveness is more evident in the temporal sequence with less flickering artifacts. Please refer to our supplemental video.

5.4. Quick Adaptation on Unseen Subjects

As shown in Figure 7, we demonstrate the quick personalization of URHand to get a relightable and animatable
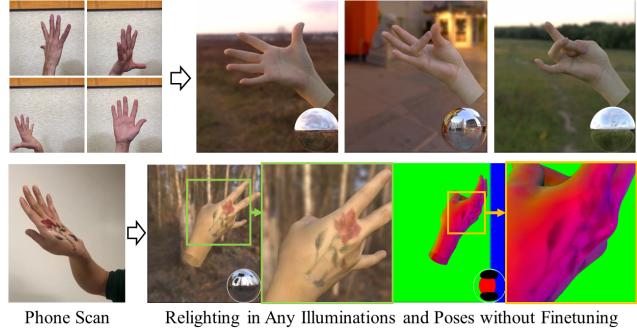

Figure 7. **Quick Personalization of URHand from iPhone captures.** Given a casual phone scan, we fit hand geometry (Sec. 3), unwrap RGB images to get the mean texture, and feed into UR-Hand. Our model instantly enables photorealistic relighting in any poses and illuminations without finetuning, even for out-of-distribution appearance (w/ tattoos) and extreme novel lighting (RGB panel). Please refer to the supplementary for more results.

hand from a monocular iPhone video. To achieve this, we followed the authentic hand avatar creation pipeline [30]. Specifically, we fit a template hand model (Sec. 3) to input images by optimizing pose parameters and the identity latent code based on foreground segmentation and 2D keypoints. Then we unwrap input RGB images based on the fitted geometry to obtain the mean texture $\mathcal{T}$, and refine it to remove shadows. Finally, we directly feed the mean texture and the coarse mesh into URHand to render it with any poses and environment maps. Our model can even generalize to out-of-distribution appearances like tattoos and challenging unseen illuminations like RGB panels (bottom row of Figure 7). Note that while URHand can instantly create a personalized relightable hand model without any finetuning, the aforementioned preprocessing stage takes several hours. Please refer to Sec. 3 of the supplementary for more implementation details and results.

6. Conclusion

We have introduced URHand, the first universal relightable hand model that generalizes across viewpoints, poses, illuminations, and identities. We show that scalable cross-identity training for high-fidelity relightable hands now is possible with our physics-inspired spatially varying linear lighting model and hybrid neural-physical learning framework. Our experiments indicate that URHand even generalizes beyond studio data by showing quick personalization from a phone scan.

References

[1] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 640–653. Springer, 2012. 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, 3

[3] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *Acm Siggraph*, pages 1–7. vol. 2012, 2012. 2, 4

[4] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8683–8693, 2023. 2

[5] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *European Conference on Computer Vision*, pages 606–623. Springer, 2022. 2, 3

[6] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20533–20543, 2022. 2

[7] Martin de La Gorce, David J Fleet, and Nikos Paragios. Model-based 3d hand pose estimation from monocular video. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1793–1805, 2011. 2

[8] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 2, 3

[9] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. Dart: Articulated hand model with diverse accessories and rich textures. *Advances in Neural Information Processing Systems*, 35:37055–37067, 2022. 2

[10] 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. 3

[11] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14719–14728, 2021. 3

[12] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4226, 2022. 3

[13] Umar Iqbal, Akin Caliskan, Koki Nagano, Sameh Khamis, Pavlo Molchanov, and Jan Kautz. Rana: Relightable articulated neural avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23142–23153, 2023. 3

[14] Shun Iwase, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Bagautdinov Timur, Rohan Joshi, Fabian Prada, Takaaki Shiratori, Yaser Sheikh, and Jason Saragih. Relightablehands: Efficient neural relighting of articulated hand models. In *CVPR*, 2023. 2, 3, 5, 6, 7

[15] 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

[16] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2023. 3

[17] Yoshihiro Kanamori and Yuki Endo. Relighting humans: occlusion-aware inverse rendering for full-body human images. *arXiv preprint arXiv:1908.02714*, 2019. 3

[18] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *2021 International Conference on 3D Vision (3DV)*, pages 11–21. IEEE, 2021. 2

[19] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. Harp: Personalized hand reconstruction from a monocular rgb video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12802–12813, 2023. 3, 6

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] 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

[22] Junxuan Li, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Hongdong Li, and Jason Saragih. Megane: Morphable eyeglass and avatar network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12769–12779, 2023. 3

[23] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019. 3

[24] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 2, 6

[25] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 5

[26] 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. 3

[27] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 2

[28] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018. 2

[29] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 440–455. Springer, 2020. 2

[30] Gyeongsik Moon, Weipeng Xu, Rohan Joshi, Chenglei Wu, and Takaaki Shiratori. Authentic hand avatar from a phone scan via universal hand model. In *CVPR*, 2024. 3, 8

[31] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1154–1163, 2017. 2

[32] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–59, 2018.

[33] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (ToG)*, 38(4):1–13, 2019. 2

[34] Akshay Mundra, Mallikarjun B R, Jiayi Wang, Marc Habermann, Christian Theobalt, and Mohamed Elgharib. Livehand: Real-time and photorealistic neural hand rendering, 2023. 2

[35] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020. 3

[36] 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. 3, 5, 7

[37] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4670–4680, 2023. 2, 6

[38] Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 54–71. Springer, 2020. 2

[39] Haonan Qiu, Zhaoxi Chen, Yuming Jiang, Hang Zhou, Xiangyu Fan, Lei Yang, Wayne Wu, and Ziwei Liu. Relitalk: Relightable talking portrait generation from a single video. *arXiv preprint arXiv:2309.02434*, 2023. 3

[40] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017. 2

[41] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018. 3

[42] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2

[43] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE international conference on computer vision*, pages 2456–2463, 2013. 2

[44] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2015. 2

[45] 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. 3

[46] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[47] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, pages 703–735. Wiley Online Library, 2022. 2

[48] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM Transactions on Graphics (ToG)*, 35(6):1–11, 2016. 2

[49] Bohan Wang, George Matcuk, and Jernej Barbič. Hand modeling and simulation using stabilized magnetic resonance

imaging. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2

[50] Yifan Wang, Aleksander Holynski, Xiuming Zhang, and Xuaner Zhang. Sunstage: Portrait reconstruction and relighting using the sun as a light stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20792–20802, 2023. 3

[51] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 5

[52] 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. 3

[53] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, et al. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (ToG)*, 25(3):1013–1024, 2006. 2

[54] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, pages 641–676. Wiley Online Library, 2022. 2

[55] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (ToG)*, 37 (4):1–13, 2018. 3

[56] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3

[57] Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. Towards practical capture of high-fidelity relightable avatars. In *SIGGRAPH Asia 2023 Conference Proceedings*, 2023. 3, 5, 6, 7

[58] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 41(6): 1–21, 2022. 3

[59] Wang Yifan, Lukas Rahmann, and Olga Sorkine-hornung. Geometry-consistent neural shape representation with implicit displacement fields. In *International Conference on Learning Representations*, 2022. 4

[60] 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

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[62] 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. 3

[63] 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. 2

[64] Mianlun Zheng, Bohan Wang, Jingtao Huang, and Jernej Barbič. Simulation of hand anatomy using medical imaging. *ACM Transactions on Graphics (TOG)*, 41(6):1–20, 2022. 2