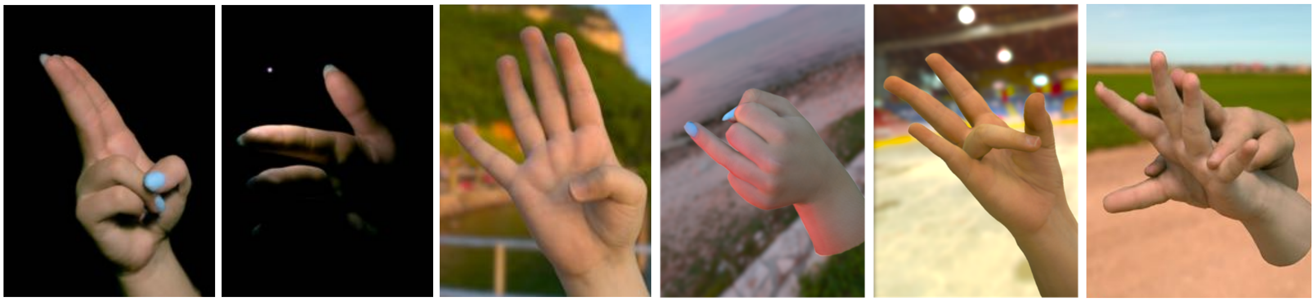
RelightableHands: Efficient Neural Relighting of Articulated Hand Models

Shun Iwase[1,2*] Shunsuke Saito[2] Tomas Simon[2]
Stephen Lombardi[2] Timur Bagautdinov[2] Rohan Joshi[2]
Fabian Prada[2] Takaaki Shiratori[2] Yaser Sheikh[2] Jason Saragih[2]
[1]Carnegie Mellon University [2]Reality Labs Research



Directional Light | Near-field Light | Environment Map | Student Model Rendering (21ms) | Two Hand Rendering (32ms)

Figure 1. **Neural Relighting of Animatable Hands.** Our model-based neural rendering approach enables high-fidelity rendering of hands with various poses, views, and illuminations. Our student model is highly efficient enough to render in real-time.

Abstract

We present the first neural relighting approach for rendering high-fidelity personalized hands that can be animated in real-time under novel illumination. Our approach adopts a teacher-student framework, where the teacher learns appearance under a single point light from images captured in a light-stage, allowing us to synthesize hands in arbitrary illuminations but with heavy compute. Using images rendered by the teacher model as training data, an efficient student model directly predicts appearance under natural illuminations in real-time. To achieve generalization, we condition the student model with physics-inspired illumination features such as visibility, diffuse shading, and specular reflections computed on a coarse proxy geometry, maintaining a small computational overhead. Our key insight is that these features have strong correlation with subsequent global light transport effects, which proves sufficient as conditioning data for the neural relighting network. Moreover, in contrast to bottleneck illumination conditioning, these features are spatially aligned based on underlying geometry, leading to better generalization to unseen illuminations and poses. In our experiments, we demonstrate the efficacy of our illumination feature representations, outperforming baseline approaches. We also show that our approach can photorealistically relight two interacting hands at real-time speeds. https://sh8.io/#/relightable_hands

1. Introduction

Neural rendering approaches have significantly advanced photorealistic face rendering [42, 55, 66] in recent years. These methods use deep neural networks to model the light transport on human skin [11, 14, 31, 63], directly reproducing physical effects such as subsurface scattering by reconstructing real images. However, despite the success of neural relighting, extending this approach to animatable hand models poses a unique challenge: generalization across articulations.

Unlike faces, hands have many joints, and the state of a single joint affects all child joints. This leads to extremely diverse shape variations even within a single subject. Changes in pose drastically affect the appearance of hands, creating wrinkles, casting shadows, and interreflecting across topologically distant regions. Rendering these effects is challenging because sufficiently accurate geometry and material properties required for photorealism are difficult to obtain, and even then, path tracing to sufficient accuracy is computationally expensive. The use of simplified geometric and appearance models (such as linear blend skinning and reduced material models) allow faster computation but come at a noticeable degradation in rendering fidelity. So far, photorealistic rendering of animatable

[*] This work was done during an internship at Meta

hands with global illumination effects in real-time remains an open problem.

In this work, we aim to enable photorealistic rendering of a personalized hand model that can be animated with novel poses, in novel lighting environments, and supports rendering two-hand interactions. To this end, we present the first neural relighting framework of a parametric 3D hand model for real-time rendering. Specifically, we build a relightable hand model to reproduce light-stage captures of dynamic hand motions.

Inspired by [4], we capture performances under spatiotemporal-multiplexed illumination patterns, where fully-on illumination is interleaved to enable tracking of the current state of hand geometry and poses. We use a two-stage teacher-student approach to learn a model that generalizes to natural illuminations outside of the capture system. We first train a teacher model that infers radiance given a point-light position, a viewing direction, and light visibility. As this model directly learns the mapping between an input light position and output radiance, it can accurately model complex reflectance and scattering on the hand without the need for path tracing. To render hands in arbitrary illuminations, we treat natural illuminations as a combination of distant point-light sources by using the linearity of light transport [9]. We then take renderings from the teacher model as pseudo ground-truth to train an efficient student model that is conditioned on the target environment maps.

However, we found that the student model architecture used in [4] for faces leads to severe overfitting when applied to relightable hands. This is caused by the architecture design of holistically conditioning a bottleneck representation with the target lighting environment. This representation makes it difficult to reproduce geometric interactions between lights and hand pose, such as those required to cast shadows from the fingers onto the palm across all possible finger configurations.

Therefore, motivated by recent neural portrait relighting works [42, 61], we instead propose to compute spatially aligned lighting information using physics-inspired illumination features, including visibility, diffuse shading, and specular reflections. Because these features are based on geometry and approximate the first bounce of light transport, they show strong correlation with the full appearance and provide sufficient conditioning information to infer accurate radiance under natural illuminations. In particular, visibility plays a key role in disentangling lights and pose, reducing the learning of spurious correlations that can be present in limited training data. However, computing visibility at full geometric resolution for every single light is too computationally expensive for real-time rendering. To address this, we propose using a coarse proxy mesh that shares the same UV parameterization as our hand model for computing the lighting features. We compute the features at vertices of the coarse geometry, and use barycentric interpolation to create texel-aligned lighting features. Our fully convolutional architecture learns to compensate for the approximate nature of the input features and infers both local and global light transport effects. This way, our model can render appearance under natural illuminations at real-time framerates as shown in Figure 1.

Our study shows that both integrating visibility information and spatially aligned illumination features are important for generalization to novel illuminations and poses. We also demonstrate that our approach supports rendering of two hands in real-time, with realistic shadows cast across hands.

Our contributions can be summarized as follows:

- The first method to learn a relightable personalized hand model from multi-view light-stage data that supports high-fidelity relighting under novel lighting environments.
- An illumination representation for parametric model relighting that is spatially aligned, leading to significant improvements in generalization and accuracy of shadows under articulation.
- An efficient algorithm to compute spatially-aligned lighting features with visibility and shading information incorporated using a coarse proxy mesh, enabling real-time synthesis.

2. Related Work

In the following, we review image-space and model-based relighting approaches as well as hand modeling techniques.

Hand Modeling Modeling human hands has been extensively studied in both computer vision and graphics. Early work primarily focuses on tracking geometry and modeling articulation. Various hand shape representations have been proposed including simple shape primitives [41, 46, 57], sum of 3D Gaussians [53, 54], sphere mesh [58], and triangle meshes [3, 8, 48, 59]. MANO [48] presents a parametric mesh model that learns identity variations as well as pose dependent deformations. Facilitated by such parametric models and accurate joint detection methods [51], estimating 3D hand poses is now possible from RGB-D inputs [33, 38] or images [34, 36, 69]. They are also extended to two hands [25, 37] and object interactions [16]. While these approaches show impressive robustness, geometric fidelity remains limited. To further improve fidelity of geometry modeling, anatomical priors from medical images [27, 60, 68] and physics-based volumetric prior [52] allows modeling more accurate surface deformation, especially around articulation and contacts. Self-supervised learning enables the learning of more personalized articulated models in an end-to-end manner with a mesh representation [35] and neural fields [1, 10, 21, 49].

The appearance of hands is also essential for realistic animation. HTML [45] builds a database of hand textures to create a parameteric texture space that can be fit to novel hands. Neural rendering approaches based on volumetric representation have been extended to articulation modeling, compensating for inaccurate geometry by using view-dependent appearance [40, 44]. In particular, LISA [7] demonstrates the modeling of animatable hands from multi-view images. However, these approaches pre-integrate illuminations into the appearance model, and relighting is not supported. NIMBLE [27] captures diffuse, normal, and specular maps from a light-stage and build a PCA appearance space. While reflactance maps allow relighting, physically-based rendering requires expensive ray-tracing and is sensitive to geometry quality, while linear appearance models have limited capacity for compensating geometry errors.

In contrast, our work proposes an end-to-end model for geometry and relightable appearance by leveraging neural rendering techniques. By directly reproducing complex light transport effects using neural rendering, our method can achieve significantly more efficient photorealistic relighting.

Image-space Human Relighting Image-space relighting has been pioneered by Devebec *et al*. [9], where faces under novel illuminations are generated by making use of the linearity of light transport from a one-light-at-a-time (OLAT) capture. A follow-up work by Wenger *et al*. [62] enables dynamic relighting by warping adjacent frames with time-multiplexed illumination patterns. The learning-based approach of Xu *et al*. [64] proposes to interpolate light positions from sparse observations, and a similar approach is extended to light-stage captures by upsampling light directions [56]. Meka *et al*. [32] also infer OLAT images from a pair of spherical gradient illuminations, enabling dynamic captures. In contrast to these approaches based on single point lights, Sun *et al*. [55] directly regress faces under natural illuminations using deep neural networks. A parallel line of work aims to decompose images into geometry and reflectances, enabling physically based relighting [17, 18, 20, 23, 39, 50]. Recent works leverage the best of learning-based relighting and material decomposition by feeding physics-inspired relighting results from the estimation into another network to produce a final relighting image [19, 42, 61, 66]. Despite plausible relighting results, image-space approaches typically suffer from temporal and view inconsistency artifacts during animation or novel-view rendering, as they lack a 3D parameterization of the scene.

Model-based Human Relighting In contrast to image-space neural relighting approaches, we can leverage a 3D template-model for animation and novel-view rendering. Yamaguchi *et al*. [65] infer skin reflectance from a single image in a shared UV space, allowing them to relight faces from different views. Zhang *et al*. [67] also leverage a shared UV space to relight novel-views of human performance captures with global light transport. Unfortunately, this approach only supports a playback of existing performances and cannot create new animations. To enable animatable relighting for facial performance, Bi *et al*. [4] presents DRAM, a deep relightable appearance model that is conditioned on viewing direction and expression latent codes. While their approach enables efficient relighting for real-time animation using a teacher-student framework, we observe that the bottleneck lighting encoding without visibility information in their student model leads to severe overfitting when applied to hand relighting. EyeNeRF [24] enables the joint learning of geometry and relightable appearance of a moving eyeball model. Compared to eyes, hands exhibit significantly more diverse pose variations, making explicit visibility incorporation essential. Relighting4D [6] learns relightable materials of an articulated human under a single unknown illumination, but the fidelity of relighting is limited bu the expressiveness of their parametric BRDF model. In contrast to these methods, our approach enables relighting of articulate hand models that can be animated with a wide range of poses. In addition, the proposed lighting encoding makes our relightable model generalizable to novel poses and illuminations while retaining real-time performance.

3. Preliminaries

Data Acquisition. We use a multiview calibrated capture system consisting of 106 cameras and 460 white LED lights to capture both fully-lit and partially-lit images of hands in motion, using a setup similar to [4]. Images are captured at 4096×2668 resolution at 90 frames per second. We represent the state of a hand using pose parameters, and estimate them for all the frames in the following way: for fully-lit frames, we perform skeletal hand tracking using a personalized Linear Blend Skinning (LBS) model. Specifically, we first obtain 3D reconstruction meshes using [15] and detect 3D hand keypoints using [26] with a ResNet backbone and RANSAC-based triangulation. An LBS model is personalized using reconstructions and keypoints on a collection of key frames, and is used for skeletal tracking [12] to estimate pose parameters for fully-lit frames. For partially-lit frames, we perform spherical linear interpolation of the pose parameters from adjacent fully-lit frames. Our dataset contains independently captured sequences of right and left hands. We collected $92,313$ and $88,413$ frames for Subject 1's hands and $22,754$ and $22,354$ frames for Subject 2 from 106 cameras. 80% of the segments are used for training and the rest for testing.

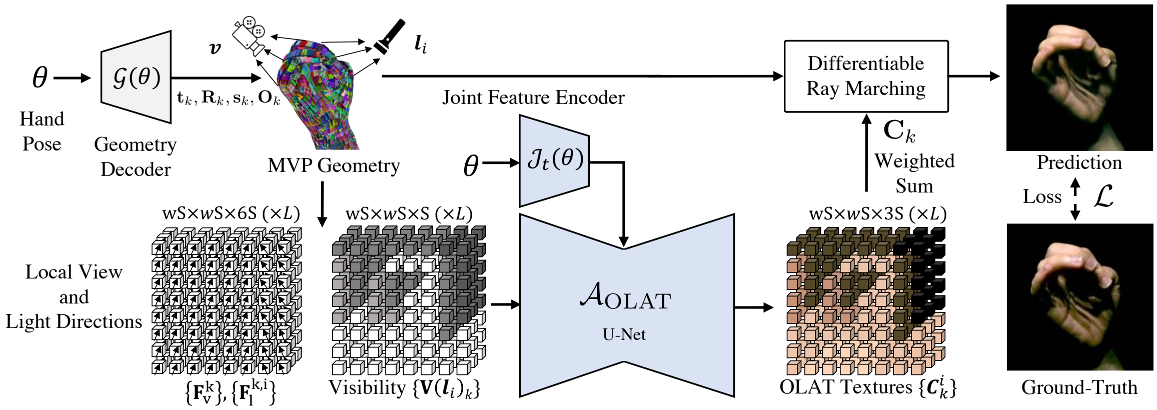

θ
Hand Pose
$\mathcal{G}(\theta)$
Geometry Decoder
v
$\mathbf{t}_k, \mathbf{R}_k, \mathbf{s}_k, \mathbf{O}_k$
l_i
MVP Geometry
Joint Feature Encoder
Differentiable Ray Marching
Prediction
Loss
$\mathcal{L}$
Ground-Truth
$\mathbf{C}_k$
Weighted Sum
θ
$\mathcal{J}_t(\theta)$
$\mathcal{A}_{\mathrm{OLAT}}$
U-Net
wS×wS×6S (×L)
wS×wS×S (×L)
wS×wS×3S (×L)
Local View and Light Directions
$\{\mathbf{F}_v^k\}, \{\mathbf{F}_l^{k,i}\}$
Visibility $\{\mathbf{V}(l_i)_k\}$
OLAT Textures $\{\mathcal{C}_k^i\}$

Figure 2. **Overview of the teacher model.** Our texture decoder U-Net takes as input view and light directions in local primitive coordinates, and visibility at each voxel in the primitives. Input pose parameters are encoded into joint features, and fed into the bottleneck layer of U-Net. The output OLAT textures are aggregated by the weighted sum using the intensity of each light. The network weights are trained in an end-to-end manner via inverse rendering losses.

Articulated Geometry Modeling. We adopt the articulated mixture of volumetric primitives (MVP) [47], which extends the original work of MVP [30] to articulated objects. As demonstrated in [47], the articulated MVP improves fidelity over mesh-based representations due to its volumetric nature while being computationally efficient for real-time rendering. Additionally, it only requires a coarse mesh from an LBS model as guidance, in contrast to prior mesh-based works [2, 29], which rely on precise surface tracking.

Given a coarse articulated mesh $\mathcal{M} = \{\mathcal{V}, \mathcal{T}, \mathcal{F}, \theta\}$ with vertices $\mathcal{V}$, texture coordinates $\mathcal{T}$, faces $\mathcal{F}$, and hand pose parameters θ representing joint angles, we decode a set of volumetric primitives. Specifically, our pose-dependent hand geometry is modeled by N primitives, where the k-th primitive is defined by $\mathcal{P}_k = \{\mathbf{t}_k, \mathbf{R}_k, \mathbf{s}_k, \mathbf{C}_k, \mathbf{O}_k\}$, comprising the primitive center location $\mathbf{t}_k \in \mathbb{R}^3$, rotation $\mathbf{R}_k \in SO(3)$, per-axis scale $s_k \in \mathbb{R}^3$, and voxels that contain color $\mathbf{C}_k \in \mathbb{R}^{3\times S\times S\times S}$ and opacity $\mathbf{O}_k \in \mathbb{R}^{S\times S\times S}$ for each primitive, where S denotes the resolution of voxels on each axis. To explicitly model articulations, primitives are loosely attached to the articulated mesh $\mathcal{M}$ produced by LBS. Given pose θ, the geometry decoder $\mathcal{G}(\theta)$ predicts residual rotations, translations, and scale together with the opacity of primitives $\{\mathbf{O}_k\}_{k=1}^N$. The texture decoder $\mathcal{C}(\theta)$ predicts the color of primitives $\{\mathbf{C}_k\}_{k=1}^N$. Both color and opacity decoders employ a sequence of 2D transpose convolutions, and the channel dimension in the last layer additionally stacks the depth-axis of each primitive's voxels. The decoded primitives $\{\mathcal{P}_k\}_{k=1}^N$ are rendered using differentiable ray marching [30]; we refer to [47] for details.

In this work, we first train the articulated MVP [47] from fully-lit images without relighting to obtain a personalized geometry decoder $\mathcal{G}(\theta)$. After training, we discard the non-relightable texture decoder $\mathcal{C}(\theta)$ and learn a relightable appearance decoder.

4. Method

Our goal is to build a relightable appearance model for hands that can be rendered under natural illuminations in real-time from a light-stage capture based on point lights. To this end, we use a similar teacher-student framework as proposed in [4], but extend it to articulated MVPs. The teacher model learns OLAT relightable textures using the partially-lit frames. Because the teacher model computes illumination for single point light sources, it generalizes to arbitrary illuminations due to the linearity of light transport [9]. However, multiple OLAT textures need to be generated to obtain a rendering under natural illumination. This leads to significant computational overheads for rendering (∼30s per frame). Thus, we use the teacher OLAT model to synthesize images under natural illuminations, and train an efficient student model that can be conditioned by an environment map to match with the pseudo ground-truth generated by the teacher model.

4.1. Teacher Model

Our teacher model $\mathcal{A}_{\mathrm{OLAT}}$ predicts the appearance of the hand model under OLAT as follows:

$$\{\mathbf{C}_k^i\}_k = \mathcal{A}_{\mathrm{OLAT}}(\theta, \mathbf{v}, \mathbf{l}_i, \{\mathbf{V}(\mathbf{l}_i)_k\}_k), \tag{1}$$

where $\mathbf{v}$ is the viewer's position, $\mathbf{l}_i$ is the position of i-th point light, and $\mathbf{V}(\mathbf{l}_i)_k \in \mathbb{R}^{S\times S\times S}$ are the visibility maps from the primitive to light $\mathbf{l}_i$ computed using Deep Shadow Maps [28]. Instead of OLAT, our partially lit frames use L=5 grouped lights to increase brightness and reduce motion blur. By leveraging the linearity of light transport,

Raymarching from each light to primitives and accumulating opacity.

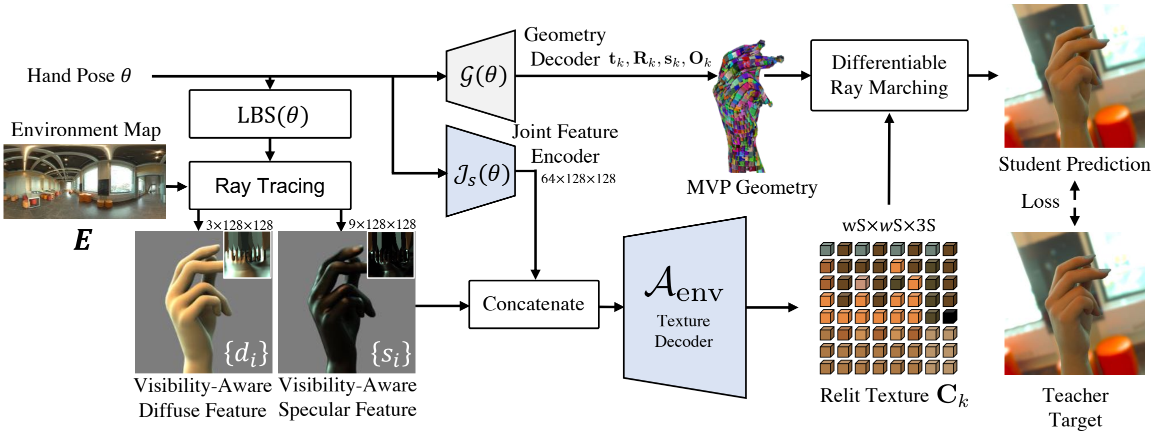


Figure 3. **Overview of the student model.** Given a hand pose and a target envmap, the visibility-aware diffuse and specular features are computed on the coarse LBS mesh. These features are then projected onto UV map and fed into the texture decoder together with the joint encoding. Finally, the predicted texture is rendered to image space via differentiable ray marching for supervision.

the final color $\{\mathbf{C}_k^i\}$ under these lights is computed as the weighted sum of primitive colors for each light:

$$\mathbf{C}_k = \sum_{i=1}^{L} b_i \mathbf{C}_k^i, \tag{2}$$

where b_i is the intensity of each light and L is the total number of lights.

Compared to a mesh-based OLAT teacher model for face relighting [4], we made several important modifications in the architecture design to support hand relighting with a hybrid mesh-volumetric representation. In [4], view-dependent intrinsic feature maps are generated at 512×512 resolution, and per-texel features are further transformed by an MLP together with the incoming light direction to infer radiance. However, an MVP-based decoder requires significantly larger channel dimensions to represent the additional volumetric depth axis. Hence, decoding radiance at every single voxel using an MLP is not computationally tractable. To address this, we adopt a U-Net architecture that takes as input reshaped visibility maps and spatially aligned light and view directions. Namely, for each primitive k and light i, light directions are encoded as $\mathbf{F}_l^{k,i} \in \mathbb{R}^{3\times S\times S\times S}$ and viewing directions as $\mathbf{F}_v^k \in \mathbb{R}^{3\times S\times S\times S}$. These are arranged into UV space as in [30, 47] to produce volumetric texture maps of size $\{\mathbf{F}_v^k\}_{k=1}^N \in \mathbb{R}^{3\times wS\times wS\times S}$, with w=64 the number of primitives per side of the UV map layout, and $N = w\times w$ the total number of primitives, and similarly for light directions. Figure 2 illustrates the overall architecture of the teacher model.

While the spatially aligned light directions are computed in a model-centric space in [4], this global parameterization leads to severe overfitting for articulated objects because it ignores local orientation changes produced by articulation. To address this, we propose to reorient view and light directions into primitive-centric coordinates. More specifically, the view directions $\mathbf{F}_v^k$ at each primitive k are represented as follows:

$$\left[\mathbf{F}_v^k\right]_j = \mathbf{R}_k^\intercal \left(\mathbf{v} - \mathbf{p}_{k,j}\right) ||\mathbf{v} - \mathbf{p}_{k,j}||_2^{-1}, \tag{3}$$

where $[\cdot]_j$ indexes voxels inside the primitive, $\mathbf{R}_k^\intercal$ is the inverse rotation matrix of the k-th primitive and $\mathbf{p}_{k,j}$ denotes the 3D location of the j-th voxel inside the k-th primitive. Similarly, the light directions $\mathbf{F}_l^{k,i}$ are expressed as follows:

$$\left[\mathbf{F}_l^{k,i}\right]_j = \mathbf{R}_k^\intercal \left(\mathbf{l}_i - \mathbf{p}_{k,j}\right) ||\mathbf{l}_i - \mathbf{p}_{k,j}||_2^{-1}, \tag{4}$$

where $\mathbf{l}$ is the location of the point light.

Additionally, joint features are input at the lowest resolution level of the U-Net layer such that the resulting appearance explicitly accounts for pose-dependent texture changes, such as small wrinkles, that may not be represented by the primitive geometry. We use a spatially aligned joint feature encoder $\mathcal{J}_t(\theta)\in\mathbb{R}^{64\times 64\times 64}$ in UV space as in [2]. Our loss is expressed by

$$\mathcal{L} = \lambda_{MSE}\mathcal{L}_{MSE} + \lambda_{VGG}\mathcal{L}_{VGG} + \lambda_{neg}\mathcal{L}_{neg}, \tag{5}$$

where $\mathcal{L}_{MSE}$ is the mean-squared error between ground-truth and rendered images, $\mathcal{L}_{VGG}$ is the weighted sum of the VGG feature loss at each layer, $\mathcal{L}_{neg}$ is a regularization term and penalizes the texture with a negative intensity:

$$\mathcal{L}_{neg} = \frac{\gamma}{NS^3} \sum_{k}^{N} ||\max(-\mathbf{C}_k, 0)||_2^2, \tag{6}$$

with weight schedule $\gamma = \exp\left(-\max\left(\eta_{neg}\frac{t}{t-t_s}, 0\right)\right)$ where t is the number of current iterations, and t_s is the iteration when the regularization loss starts decaying.

4.2. Student Model

Our student model $\mathcal{A}_{\text{env}}$ predicts the appearance of the hand model under natural illuminations represented as en-

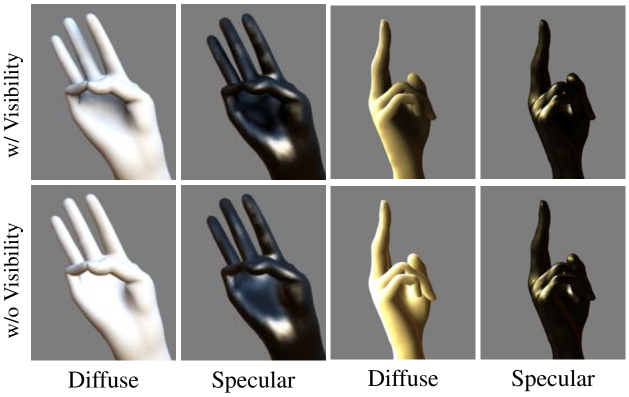


Figure 4. **Effect of visibility integration.** The visibility integration based on ray tracing leads to accurate encoding of shadow information in both specular and diffuse features.

vironment maps $\mathbf{E} \in \mathbb{R}^{M\times 3}$ as follows:

$$\{\mathbf{C}_k\} = \mathcal{A}_{\text{env}}(\theta, \mathbf{E}). \tag{7}$$

The principal challenge in the student model is efficient light encoding that can be generalized to illuminations and poses unseen during training. The state-of-the-art model-based approach for face relighting [4] introduces an efficient hypernet architecture conditioned on a 512-dimensional bottleneck representation of a 16 × 32 environment map. We find that such a bottleneck illumination representation is difficult to generalize because it loses the spatial geometry of illumination by collapsing it into a single vector. As we demonstrate in our evaluation (Sec. 5.2), this issue is even more pronounced for relightable hands, as it is non-trivial to disentangle global light transport between such bottleneck illumination and pose features. A similar observation has been made for portrait relighting, where methods based on image-aligned light features [42, 66] outperform an approach using a bottleneck representation [55]. Inspired by this success, we propose a spatially aligned illumination representation tailored for model-based hand relighting that accurately accounts for self-occlusion due to pose changes.

Namely, we produce our texel-aligned feature representation by casting M rays (one per envmap location) from each vertex and compute a weighted sum of envmap values to produce diffuse and specular components. We incorporate the visibility information by setting the contribution of rays hitting other mesh parts to zero. Figure 4 shows the effect of the visibility integration. In practice, we use a coarse mesh to compute per-vertex features, which are then projected to texel-aligned space via barycentric interpolation. The projected features are fed into a fully convolutional decoder, retaining the spatial alignment between features and the output appearance $\{\mathbf{C}_k\}$. Figure 3 illustrates the overall architecture of the student model.

More precisely, the diffuse feature $\mathbf{d}_i \in \mathbb{R}^3$ at vertex i is represented with Lambertian BRDF computed as follows:

$$\mathbf{d}_i = \sum_{m=1}^{M} \mathbf{E}(\mathbf{r}_i^m) h_i(\mathbf{r}_i^m) \max(\mathbf{n}_i \cdot \mathbf{r}_i^m, 0) \ , \tag{8}$$

where $\mathbf{E}(\mathbf{r}_i^m) \in \mathbb{R}^3$ is the vector of the envmap intensity sampled along the ray direction $\mathbf{r}_i^m \in \mathbb{R}^3$ (based on a far-field environment assumption), $h_i(\mathbf{r}_i^m)$ is a binary visibility term, and $\mathbf{n}_i \in \mathbb{R}^3$ is a vertex normal. The specular feature $\mathbf{s}_i(\alpha) \in \mathbb{R}^3$ is represented with Phong specular BRDF computed as

$$\mathbf{s}_i(\alpha) = \sum_{m=1}^{M} \mathbf{E}(\mathbf{r}_i^m) h_i(\mathbf{r}_i^m) \max(\hat{\mathbf{v}}_i \cdot \mathbf{r}_i^m, 0)^\alpha \ , \tag{9}$$

where α is a shininess coefficient, and $\hat{\mathbf{v}}_i$ is the view direction reflected around the normal. To account for spatially varying material properties on hands, we take specular features with multiple shininess values $(16, 32, 64)$. The feature maps are also concatenated with spatially aligned joint features $\mathcal{J}_s(\theta) \in \mathbb{R}^{64\times 128\times 128}$.

Note that we train the student model $\mathcal{A}_{\text{env}}$ using the same losses used for the teacher model (Eq.5).

4.3. Implementation Details

To train the teacher and student models, we use Adam [22] optimizer and set the hyperparameters λ_{MSE}, λ_{VGG}, and λ_{neg} to 1.0, 1.0, and 0.01 respectively. We train each of the geometry module, teacher model, and student model for $100,000$ iterations with the learning rate of 0.001, and batch size of 4, 2, and 4, respectively, on NVIDIA V100 and A100. In addition, we use N=4096 primitives whose per-axis resolution S is 16. We describe the detailed network architecture in the supplemental. To reduce redundancy in our training data in terms of poses, we adopt importance sampling based on kernel density estimation using the subset of tracked hand vertices in root-normalized coordinates. We generate $25,000$ images with 1000 frames and 25 cameras to train the student model. To compute the texel-aligned lighting features, we use envmap of size M=512(16×32), and coarse mesh with 2825 vertices. The visibility is computed with a GPU-accelerated triangle-ray intersection using NVIDIA OptiX [43].

5. Experimental Results

We evaluate our method using 2 subjects with left, right and two hands captured in a light-stage as described in Sec. 3. For evaluation, we exclude several segments to assess the generalization of our model to novel poses. To evaluate the generalization of the student model to novel illumination, we use 3094 high-resolution HDR environment maps consisting of the ones in [13] and [55]. We use 2560

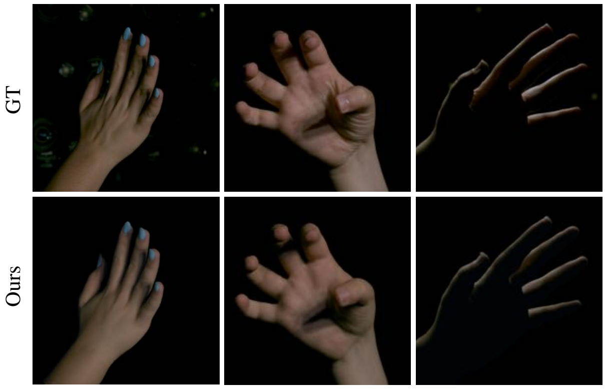


Figure 5. **Qualitative results of the teacher model.** We evaluate our method using the ground-truth images captured in the light-stage. Our approach successfully models inter-reflection, shadow, and subsurface scattering.

of them for training and 534 for evaluation. Note that the two-hand sequences are used only for teacher model evaluation. We also compare our approach with the state-of-the-art model-based relighting method [4]. As [4] was originally proposed for face relighting, we make several modifications for fair comparison. Please refer to the supplemental for details.

We report mean squared error (MSE) and Structural Similarity Index (SSIM) to measure the quality of the generated images by the teacher and student models. To solely evaluate the quality of hands, we remove the background by using a mask image obtained from the tracked hand geometry.

5.1. Evaluation of Teacher Models

We first evaluate the effectiveness of the proposed teacher model using the images captured with the light-stage as ground-truth.

Qualitative Evaluation. We evaluate the quality of the images generated by our model against the real images captured in the light-stage. As shown in Figure 5, our teacher model is able to reproduce diverse pose-dependent appearance under multiple point-light sources, such as shadows on the wrinkles and reflection on the skin and nails.

Ablation on Visibility Conditioning. Table 1 shows that the visibility input significantly improves the accuracy on all the metrics especially for the two hand sequences. As shown in Figure 6, while the model without visibility input can overfit to training poses by relying on the joint information, it does not generalize to unseen poses or two hand cases. Thus, the visibility conditioning is essential to generalize beyond training pose and light distributions.

5.2. Evaluation of Student Models

Table 2 shows that our method achieves the best MSE and SSIM scores on all the metrics under unseen hand poses and illumination settings. Since we do not have ground-truth real images under natural illumination, we use test images generated by the teacher model for evaluation. To confirm the effectiveness of our method, we compare our method against the latest model-based relighting method [4] and perform ablation study on visibility awareness and specular features.

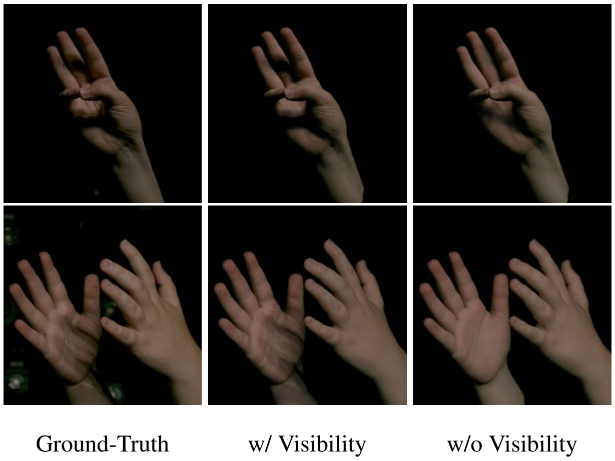


Figure 6. **Ablation on visibility conditioning of the teacher model.** The lack of visibility features leads to incorrect shadows.

Comparison to Bottleneck Light Conditioning. To evaluate the effectiveness of our spatially aligned lighting features, we compare against a bottleneck light representation with a hyper-network proposed in [4]. For fair comparison, we replace our illumination encoding with their hyper-network while retaining everything else. Figure 7 shows that a bottleneck-based light encoding fails to match the overall intensity compared to the ground-truth. Moreover, it lacks fine-grained illumination effects such as reflection from grazing angles and soft shadows. In contrast, our spatially aligned representation even without the proposed visibility integration significantly improves the fidelity of reconstruction. This observation is also strongly supported by our quantitative evaluation as shown in Table 2.

Ablation on Visibility Integration. We also evaluate the effectiveness of the proposed efficient visibility integration for computing features. Figure 7 and Table 2 show that, compared to the model without the visibility integration, our full model achieves more faithful reconstruction of shadows even for novel poses and illuminations.

Ablation on Specular Features. We also validate the importance of specular features in the student model. Figure 7 and Table 2 illustrate that our specular feature provides sufficient information to reproduce specular highlight despite having the feature computed on a coarse proxy geometry. This suggests that spatially aligned lighting features are essential for achieving generalizable neural relighting.

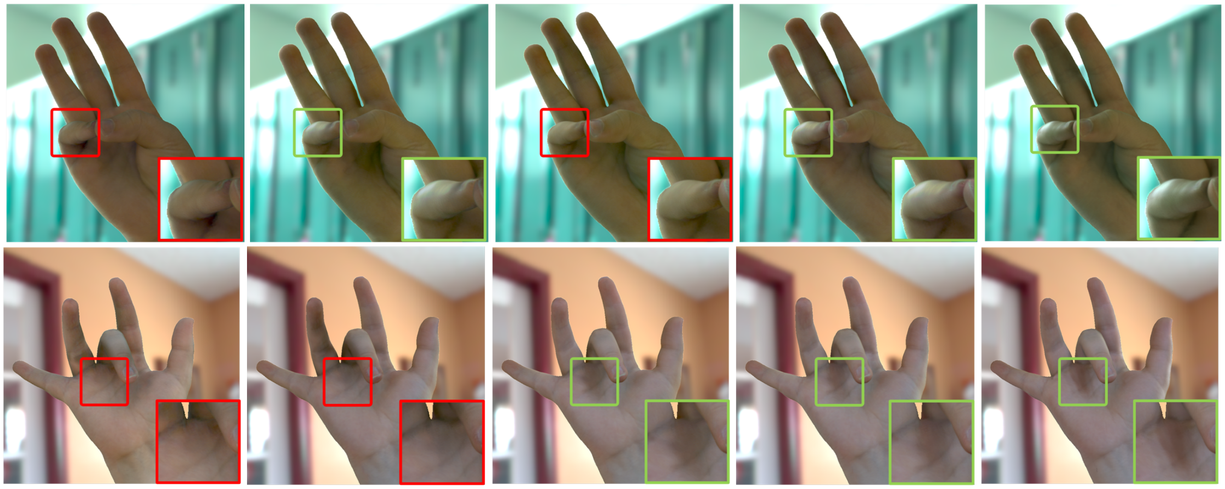
Runtime Analysis. One of our key contributions is the efficient rendering speed. While the teacher model takes ap-

	Subject 1						Subject 2					
	MSE ($\times 10^{-3}$) ↓			SSIM ↑			MSE ($\times 10^{-3}$) ↓			SSIM ↑		
	Right	Left	Both	Right	Left	Both	Right	Left	Both	Right	Left	Both
Ours	**4.9126**	**5.8608**	**15.7589**	**0.9790**	**0.9805**	**0.9536**	**8.8205**	**7.9357**	**22.3559**	**0.9541**	**0.9559**	**0.9075**
w/o Visibility	7.3201	7.8870	22.8308	0.9773	0.9792	0.9488	9.7104	9.9781	26.5647	0.9536	0.9543	0.9050

Table 1. **Quantitative comparison of the teacher model.** We measure the MSE and SSIM metrics on the right, left, and two-hand sequences. The result shows that conditioning visibility significantly improves generalization to test poses and illuminations.

	Subject 1						Subject 2					
	MSE ($\times 10^{-3}$) ↓			SSIM ↑			MSE ($\times 10^{-3}$) ↓			SSIM ↑		
	Right	Left	Both	Right	Left	Both	Right	Left	Both	Right	Left	Both
DRAM [4]	31.1372	24.4368	64.2035	0.9904	0.9927	0.9752	30.6582	24.7238	70.4215	0.9901	0.9898	0.9665
Ours	**5.4076**	**5.9600**	**4.3474**	**0.9961**	**0.9960**	**0.9915**	**5.7977**	**7.2598**	**4.5196**	**0.9952**	**0.9954**	**0.9881**
w/o Specular	5.7660	7.2631	5.0732	0.9956	0.9952	0.9914	7.1569	7.4892	4.9008	0.9948	0.9943	**0.9881**
w/o Visibility	6.6110	8.1886	11.6771	0.9955	0.9948	0.9893	7.8589	8.5550	9.1859	0.9938	0.9938	0.9862

Table 2. **Quantitative evaluation of the student model.** Our student model outperforms the state-of-the-art model-based relighting method [4] by a large margin. In addition, the proposed visibility and specular features integration significantly improve the generalization to unseen poses and natural illuminations.



DRAM [4] w/o Visibility w/o Specular Feature Ours Ground-Truth (Teacher)

Figure 7. **Qualitative comparison of the student model.** A model-based relighting method [4] fails to reproduce the precise color and fine-grained shading effects. Our model without visibility integration or specular features also lacks pose-dependent shadow or specular highlight respectively. In contrast, our full model successfully reproduces both effects. The green and red bounding boxes denote success and failure cases respectively.

proximately 30 seconds to generate a texture with an envmap by aggregating over $512 (= 16 \times 32)$ light sources, the student model achieves 48 fps (21 ms) for a single hand and 31 FPS (32 ms) for two hands on NVIDIA V100.

6. Discussion and Future Work

We introduced the first model-based neural relighting for articulated hand models to enable photorealistic rendering of personalized hands under various illuminations in real-time. We successfully extend the teacher-student framework to build articulated models using a mesh-volumetric hybrid representation from multi-view light-stage capture data. The hybrid representation allows us to use a coarse mesh to efficiently compute physics-inspired light features as input conditioning for the proposed student model. Our experiments show that the spatially aligned light representation and explicit visibility integration are critical for highly generalizable relighting to novel poses and illuminations.

Limitations and Future Work. Our student model currently does not support inter-reflection by other nearby objects due to far-field light assumption, which can be partially addressed by taking surroundings as a spatially varying envmap. Future work also includes extending the proposed approach to clothed bodies, where computing visibility at a coarse mesh would not be sufficient for recovering fine-level shading caused by clothing deformations. Another exciting direction is to build a universal relightable hand model that spans inter-subject variations. As demonstrated by recent work [5], such a universal model would enable adaptation from in-the-wild inputs.

References

[1] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *ICCV*, 2021. 3

[2] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *TOG*. 4, 5

[3] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, 2012. 2

[4] 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. *TOG*, 2021. 2, 3, 4, 5, 6, 7, 8

[5] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. Authentic volumetric avatars from a phone scan. *TOG*. 8

[6] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *ECCV*, 2022. 3

[7] 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 *CVPR*, 2022. 3

[8] Martin de La Gorce, David J Fleet, and Nikos Paragios. Model-based 3d hand pose estimation from monocular video. *PAMI*, 2011. 2

[9] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *SIGGRAPH*, 2000. 2, 3, 4

[10] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *ECCV*, 2020. 3

[11] Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul Debevec. Driving high-resolution facial scans with video performance capture. *TOG*, 2014. 1

[12] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, 2009. 3

[13] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *TOG*, 2017. 6

[14] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. *TOG*, 2011. 1

[15] 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. *TOG*, 2019. 3

[16] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2

[17] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In *CVPR*, 2022. 3

[18] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *CVPR*, 2021. 3

[19] Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. Geometry-aware single-image full-body human relighting. In *ECCV*, 2022. 3

[20] Yoshihiro Kanamori and Yuki Endo. Relighting humans: Occlusion-aware inverse rendering for full-body human images. *SIGGRAPH Asia*, 2018. 3

[21] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *3DV*, 2021. 3

[22] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[23] Manuel Lagunas, Xin Sun, Jimei Yang, Ruben Villegas, Jianming Zhang, Zhixin Shu, Belen Masia, and Diego Gutiérrez. Single-image full-body human relighting. In *Eurographics Symposium on Rendering*, 2021. 3

[24] Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C. Buehler, Otmar Hilliges, and Thabo Beeler. Eyenerf: A hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *SIGGRAPH*, 2022. 3

[25] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, 2022. 2

[26] 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:1901.00148*, 2019. 3

[27] 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. *SIGGRAPH*, 2022. 2, 3

[28] Tom Lokovic and Eric Veach. Deep shadow maps. In *SIGGRAPH*, 2000. 4

[29] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *TOG*, 2018. 4

[30] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *TOG*, 2021. 4, 5

[31] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *EGSR*, 2007. 1

[32] 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. *TOG*. 3

[33] 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 *CVPR*, 2018. 2

[34] 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 *ECCV*, 2020. 2

[35] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, 2020. 3

[36] 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 *CVPR*, 2018. 2

[37] 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. *TOG*, 2019. 2

[38] 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 *ICCV*, 2017. 2

[39] Thomas Nestmeyer, Jean-François Lalonde, Iain A. Matthews, and Andreas M. Lehrmann. Learning physics-guided face relighting under directional light. In *CVPR*, 2020. 3

[40] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021. 3

[41] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, 2011. 2

[42] 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. In *TOG*, 2021. 1, 2, 3, 6

[43] Steven G Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, et al. Optix: a general purpose ray tracing engine. *TOG*. 6

[44] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 3

[45] 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 *ECCV*, 2020. 3

[46] James M Rehg and Takeo Kanade. Visual tracking of high dof articulated structures: an application to human hand tracking. In *ECCV*, 1994. 2

[47] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, and Yaser Sheikh. Drivable volumetric avatars using texel-aligned features. 2022. 4, 5

[48] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 2017. 2

[49] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, 2021. 3

[50] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *CVPR*, 2018. 3

[51] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2

[52] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *TOG*. 2

[53] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, 2015. 2

[54] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *ICCV*, 2013. 2

[55] Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *TOG*, 2019. 1, 3, 6

[56] 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. *TOG*, 2020. 3

[57] Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust articulated-icp for real-time hand tracking. In *Computer graphics forum*, 2015. 2

[58] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *TOG*, 2016. 2

[59] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016. 2

[60] Bohan Wang, George Matcuk, and Jernej Barbič. Hand modeling and simulation using stabilized magnetic resonance imaging. *TOG*. 2

[61] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *TOG*, 2020. 2, 3

[62] Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. *TOG*, 2005. 3

[63] 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. *TOG*, 2006. 1

[64] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *TOG*, 2018. 3

[65] 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. *TOG*, 2018. 3

[66] 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. *SIGGRAPH Asia*, 2022. 1, 3, 6

[67] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, Jonathan T. Barron, Ravi Ramamoorthi, and William T. Freeman. Neural light transport for relighting and view synthesis. *SIGGRAPH*, 2021. 3

[68] Mianlun Zheng, Bohan Wang, Jingtao Huang, and Jernej Barbič. Simulation of hand anatomy using medical imaging. *SIGGRAPH Asia*, 2022. 2

[69] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, 2020. 2