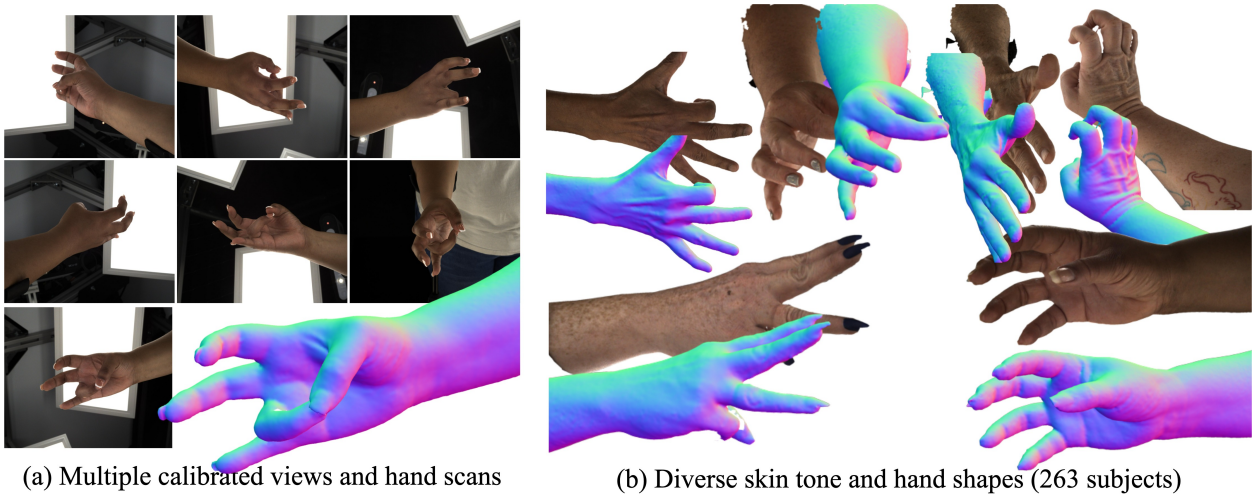
PALM: A Dataset and Baseline for Learning Multi-subject Hand Prior

Zicong Fan[1,2,3] Edoardo Remelli[1] David Dimond[1] Fadime Sener[1]
Liuhao Ge[1] Bugra Tekin[1] Cem Keskin[1] Shreyas Hampali[1]
[1]Meta Reality Labs [2]ETH Zürich [3]Max Planck Institute for Intelligent Systems, Tübingen



(a) Multiple calibrated views and hand scans (b) Diverse skin tone and hand shapes (263 subjects)

Figure 1. **Dataset overview:** PALM is a large-scale dataset comprising calibrated multi-view high-resolution RGB images and 3dMD hand scans (a). It features 263 subjects spanning a wide range of skin tones and hand sizes, $90k$ RGB images, and $13k$ high-quality hand scans with corresponding MANO registrations (b). This diversity and precision provide a foundation for learning a universal prior over human hand shape and appearance.

Abstract

The ability to grasp objects, signal with gestures, and share emotion through touch all stem from the unique capabilities of human hands. Yet creating high-quality personalized hand avatars from images remains challenging due to complex geometry, appearance, and articulation, particularly under unconstrained lighting and limited views. Progress has also been limited by the lack of datasets that jointly provide accurate 3D geometry, high-resolution multi-view imagery, and a diverse population of subjects. To address this, we present PALM, a large-scale dataset comprising $13k$ high-quality hand scans from 263 subjects and $90k$ multi-view images, capturing rich variation in skin tone, age, and geometry. To show its utility, we present a baseline PALM-Net, a multi-subject prior over hand geometry and material properties learned via physically based inverse rendering, enabling realistic, relightable single-image hand avatar personalization. PALM's scale and diversity make it a valuable real-world resource for hand modeling and related research.

1. Introduction

Human hands are central to how we interact with the physical and social world: we manipulate objects [13, 14, 19, 47, 52], express intent through gestures [11, 44, 49], and communicate affective cues via touch. Realistic and drivable hand avatars have the potential to transform virtual interaction, gaming, and telepresence. However, building such avatars from images remains a fundamentally challenging problem due to the complexity of hand geometry, appearance, and articulation, particularly under unconstrained lighting and from limited visual observations.

A critical missing component in the hand community is a large-scale, high-quality dataset that enables learning generalizable and physically grounded models of human hands. Existing datasets suffer from significant limitations: they often include only a small number of subjects [27, 30], lack

accurate 3D hand geometry from real-world scans [30, 37], or are derived from hand-crafted synthetic data [15], limiting their utility for learning models that generalize across identity and illumination.

To fill this gap, we introduce PALM, a large-scale dataset of human hands containing publicly available-ready accurate hand scans, diverse in quantity and subject diversity. PALM includes $13k$ high-quality 3D hand scans and $90k$ high-resolution multi-view RGB images from 263 subjects, each performing a diverse set of predefined hand poses designed to span a wide range of natural hand articulations. The subjects cover a broad range of skin tones and age groups. All data is captured using a commercial 3dMD scanner [43], providing precise, sub-millimeter geometry. Each scan is paired with synchronized multi-view images and a MANO registration (pose and shape), obtained via multi-view–consistent alignment to the 3D scans. Importantly, the capture environment, including lighting and scanner configuration, remained fixed throughout the entire collection process, enabling consistent illumination conditions across subjects. While prior datasets have included either limited subjects or unreleased scan data, PALM will be made publicly available for research use upon publication, making it the most comprehensive and accessible dataset for studying generalizable hand models.

To highlight a practical use case for our data, we present a baseline model PALM-Net, an implicit neural prior over human hands that jointly models appearances, geometry and material properties using our dataset. PALM-Net is trained via physically based inverse rendering, decomposing each subject hand into geometry, albedo, specularity, roughness, and environment lights. The model conditions on pose and subject-specific latent codes, enabling it to capture pose-dependent effects. A key insight in PALM-Net is a shared environment lighting constraint across the subjects, which disentangles illumination from intrinsic hand appearance, allowing the model to generalize to novel lighting.

We apply our prior to the highly under-constrained task of single-image hand avatar personalization under unknown lighting. This monocular setting presents several challenges: depth ambiguity, occlusions (*e.g.*, self-occlusion of the palm or dorsum), and ambiguous lighting. Our central insight is that, despite the fine-scale complexity of hands (*e.g.*, wrinkles, creases, texture), many fundamental properties – skin tone, material reflectance, and deformation behavior – are shared across individuals and can be captured by a learned prior. By optimizing the subject-specific latent code and scene illumination, our model reconstructs realistic, relightable, and articulated hand avatars from a single input image, even in uncontrolled conditions.

To perform extensive evaluation of the single-image personalization setting, we evaluate our method in both synthetic and real-world datasets. Our results show that our method consistently outperforms other prior-based and non-prior-based methods for relightable hand personalization.

To summarize our contributions: 1) We introduce PALM, a large-scale dataset containing high-resolution RGB multi-view images of diverse subjects with detailed hand scans and accurate MANO registrations; 2) We present a baseline, PALM-Net, a multi-subject implicit hand prior model that leverages PALM to learn physical hand properties such as geometry, albedo, specularity, and roughness; 3) We demonstrate the effectiveness of our prior model by using it to personalize and relight hand avatars from a single image under challenging and diverse environmental conditions.

2. Related Work

Hand datasets: Recent years have seen an explosion of hand datasets, that can be categorized into hand-only [15, 27, 29, 35, 59], interacting hand [30, 31, 45], and hand-object [2, 4, 12, 16, 18, 19, 23, 26, 39, 51] datasets. Early research on hand-only capture primarily focused on datasets collected using depth cameras [50]. Soon after, RGB-based datasets gained traction, with notable examples including STB [53] and FreiHAND [59]. Following this, interest expanded toward hand-object interaction datasets. Hampali *et al*. [18] released a dataset of single hands manipulating rigid YCB objects, while Fan *et al*. [12] used a marker-based MoCap setup to capture full-body interactions with articulated objects. More recently, Banerjee *et al*. [1] contributed a large-scale dataset featuring multi-object interactions. Interacting-hand datasets have also grown in popularity. Tzionas *et al*. [45] pioneered capturing two-hand interactions using an RGB-D setup. Building on this, Moon *et al*. [30] leveraged a large-scale multi-view RGB system to recover 3D hand poses under strong self-contact, which significantly boosted interest in modeling two interacting hands from RGB data. Moon *et al*. [31] released a synthetic dataset using 3D annotations derived from [30]. Handy [35] provides 3dMD data but it is not publicly available. Despite the rapid emergence of various hand datasets, most approaches for learning hand avatars still rely primarily on InterHand2.6M [7, 34] or on custom video recordings [22]. This is largely due to the absence of a high-quality dataset that simultaneously provides accurate 3D hand scans, high-resolution multi-view RGB imagery, and a diverse set of subjects.

Hand representations: Learning articulated hand representations is a long-standing research problem. Some methods solely focus on modelling hand joints [11, 41, 57] or geometries [10, 20, 29, 37, 56]. For example, MANO [37] pioneered parametric mesh-based hand geometry modeling, parameterizing shape with a PCA latent space. Moon *et al*. [29, 32] propose a non-linear approach for high-fidelity hand mesh modeling. HALO [21] is an implicit articulated hand geometry representation using occupancy network via a differentiable canonicalization layer. There are

also methods that model both geometry and appearances of hands [7–9, 14, 24, 34–36, 55]. HTML [36] extends MANO with a PCA-based texture model. Handy [35] is a parametric hand model of shape and texture learned from proprietary hand scans. Both Handy and HTML preprocess texture maps to minimize baked-in shadows and specularities. LISA and OHTA [9] model shape and appearance fields, with lighting effects baked into the network and controlled by latent codes. NIMBLE [24] models hands with bone, muscle, and skin deformation, using a light stage to capture pose-independent albedo and specular maps modeled with PCA bases. HARP [22] optimizes the normal and albedo maps for the MANO hand mesh with a point light source to model shadow effects, demonstrating slight generalizability to novel illuminations. URHand[8] models pose-dependent hand material properties and is trained on large-scale light stage data. Capturing such pose-dependent material characteristics and lighting effects requires accurate environment map information, typically enabled by a sophisticated light stage setup with hundreds of synchronized cameras, as used in Nimble and URHand. In contrast to these methods [8, 9, 55], our baseline method can jointly learn appearances, geometries and relight the hand avatar and does not rely on expensive light stage setup.

3. PALM Dataset

Overview: To study hand priors, we introduce PALM (see Figure 1), a high-quality dataset with accurate 3D hand annotations, high-resolution multi-view RGB images, and 3dMD hand scans. It contains $90k$ RGB images and $13k$ hand scans from 263 subjects (131 male, 132 female) with diverse skin tones and hand shapes. Data was captured using a 3dMD hand scanner [43] with 7 RGB and 14 monochrome machine vision cameras calibrated; the RGB images have a resolution of 2448×2048; and the 3D hand scans were reconstructed using 3dMD's software-driven triangulation technique based on Active Stereo Photogrammetry. Participants performed approximately 50 predefined right-hand gestures. See SupMat for more examples in PALM.

3.1. Data Characteristics

Dataset comparison: Table 1 compares publicly available hand-only datasets. Early datasets [33, 45, 50] focus on depth cameras and often have limited subjects and scale [33, 45]. Recent larger datasets focus on RGB images: InterHand2.6M [30] has 50 subjects with multi-view calibrated RGB, Re:InterHand [31] offers synthetically relit interacting-hand images of 10 subjects, and MANO [37] provides hand scans of 31 subjects. Handy [35] includes 3dMD scans, but these are unavailable. While MANO, InterHand2.6M, and our dataset are all suitable for learning priors for hands, most other datasets are suboptimal due to missing modalities, limited subject diversity, synthetic data, or low image quality, leading most methods [7, 22, 32, 34] to rely on InterHand2.6M despite lacking scans. Our dataset is substantially larger in scale, with 263 subjects, real high-resolution calibrated RGB images, and $13k$ high-quality scans, making it ideally suited for learning robust hand priors; it will be released publicly to advance future research.

Demographics: Figure 3 provides a detailed breakdown of

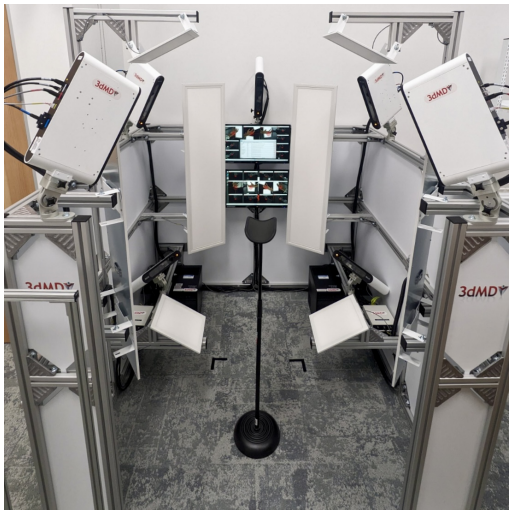
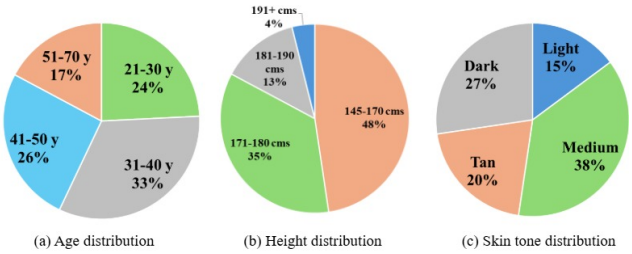


Figure 2. **Capture setup**. Our 3dMD setup with 7 RGB cameras.

dataset	# real sub.	# scans	image size	real RGB	annotation	prior learning
ICVL [42]	10	0	320 × 240	-	track	×
BigHand2.2M [50]	10	0	640 × 480	-	marker	×
Tzionas et al. [45]	-	0	640 × 480	✓	track	×
Simon et al. [40]	-	0	1920 × 1080	✓	semi-auto	×
EgoDexter [33]	4	0	640 × 480	✓	manual	×
STB [53]	1	0	640 × 480	✓	manual	×
FreiHAND [58]	32	0	224 × 224	✓	semi-auto	×
InterHand2.6M [30]	50	0	512 × 334	✓	semi-auto	✓
Re:InterHand [31]	10	0	4096 × 2668	×	-	×
DART [15]	-	0	512 × 512	×	-	×
MANO [37]	31	1K	N/A	✓	manual	✓
PALM (Ours)	263	13K	2448 × 2048	✓	semi-auto + scan	✓

Table 1. **Publicly available datasets**. Existing public datasets lack subject diversity, accurate hand scans, and high-quality multi-view RGB images that are important for training a strong hand appearance and geometry prior model. Our dataset contains a large number of subjects with high-resolution images and scans suitable for learning a universal hand prior.



(a) Age distribution (b) Height distribution (c) Skin tone distribution

Figure 3. **PALM demographics**. (a) Age; (b) Height; (c) Skin tone distributions. Our dataset provides a wide distribution of skin tones and age groups representing a large variety of hand textures.

the demographic distribution in PALM. The dataset includes participants aged 21–70 years, with the majority in the 31–40 (33%), 41–50 (26%), and 21–30 (24%) age brackets. In terms of height, subjects range from 145 to 200 cm, with 48% in the 145–170 cm range and 35% in the 171–180 cm range. Only a small portion (4%) are taller than 190 cm. Skin tone distribution is also diverse, comprising 38% medium, 27% dark, 20% tan, and 15% light tones. This diversity supports robust analysis across demographic variations.

3.2. Data Acquisition

Capture setup: We use a 3dMD [43] hand scanner to capture high-resolution multi-view RGB images and 3D scans of the subjects. In particular, the 7-viewpoint setup consists of an array of 21 synchronized and calibrated machine vision cameras, including both RGB and monochrome sensors, arranged to provide full 360-degree coverage of the hand. The system captures images at a high resolution of 2448×2048. Random light projectors are integrated to enhance surface detail and geometry accuracy. The 3D hand scans were reconstructed using 3dMD's software-driven triangulation technique based on Active Stereo Photogrammetry. All cameras are calibrated, ensuring consistent alignment across views. This setup enables the acquisition of dense, accurate 3D hand meshes in various poses, suitable for studying both static poses and dynamic hand articulations.

Capture protocol: Each participant is asked to stand still with their right hand placed inside the 3dMD capture volume. Participants are instructed to perform approximately 50 predefined right-hand gestures, covering a wide range of articulations including open-hand poses, pinches, fist closures, and fine-grained finger movements. To ensure consistency across subjects, a standardized gesture list is followed, and participants are guided through the sequence during the capture. We capture each gesture as an independent hand scan. This protocol ensures diverse and repeatable motion data across the entire subject pool. All subjects are captured in the same lighting and camera setup.

3D keypoint annotation: A semi-automatic pipeline is used to generate accurate 3D hand pose labels. The 2D keypoints are first manually annotated on a subset of images to train a 2D keypoint detector tailored to our capture setup. The detector, implemented as a U-Net [38] pre-trained on InterHand2.6M and fine-tuned on 10K manually annotated PALM images, is then used to estimate 2D keypoints for all camera views, which are subsequently triangulated to obtain 3D poses using multi-view geometry. Specifically, we follow the approach used in InterHand2.6M and apply RANSAC-based triangulation to robustly solve for the 3D keypoint locations. This semi-automatic approach significantly reduces the manual labeling burden while ensuring reliable 3D pose annotations.

MANO registration: The 3D MANO poses for each pose and subject are obtained by registering the MANO hand model to the hand scans. To register MANO to each hand gesture, we optimize the MANO hand model using a combination of 2D/3D keypoints, segmentation mask, and 3D hand scan supervision. Specifically, we minimize the closest-point distance from each MANO vertex to the corresponding hand scan surface. Ground-truth masks are derived from the hand scans, and we employ Soft Rasterizer [25] for differentiable rendering to align the MANO silhouette with these masks.

The registration is independently performed for each subject in two stages. The first stage optimizes the per-subject hand shape parameters and per-frame pose parameters on a set of simple hand poses (*e.g.*, flat hand), and the second stage only optimizes the per-frame pose parameters while freezing the shape parameters. Our final 3D keypoints have a recall rate of 95% at 10mm threshold. Our MANO registrations show a mean fitting error of 5.3 mm with respect to the 3D keypoints (similar to that of InterHand2.6M [30]).

4. Method

PALM-Net, illustrated in Figure 4, utilizes PALM, our large-scale collection of human hand data containing detailed hand scans and high-resolution RGB images, to train a *personalizable* and *relightable* hand prior. In this section, we first introduce preliminary concepts on NeRF [28], the Neural Radiance Field technique at the core of our approach, and MANO [37], the parametric hand model that we leverage within our pipeline to map 3D points across different subjects and hand poses into a shared *canonical* representation. Next, we present PALM-Net, our novel framework for learning a multi-subject hand prior through physically based inverse rendering (PBR), and describe how to train our representation over multiple subjects. Finally, we detail how PALM-Net can be used to recover a *personalized* and *relightable* hand avatar from a single image. This is a highly under-constrained problem, and we show that a prior model such as PALM-Net helps in recovering realistic, personalized hand avatars even in extreme illumination settings.

4.1. Preliminaries

NeRF: Given a ray $\mathbf{r} = (\mathbf{o}, \mathbf{d})$ defined by its camera center $\mathbf{o}$ and viewing direction $\mathbf{d}$, NeRF [28] computes the output radiance (*i.e.*, pixel color) of the ray via:

$$C_{rf}(\mathbf{r}) = \int_{t_n}^{t_f} T(t_n, t)\sigma_t(\mathbf{r}(t))L(\mathbf{r}(t), \mathbf{d})dt \qquad (1)$$
$$\text{s.t} \quad \mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$$
$$T(t_n, t) = \exp\left(-\int_{t_n}^{t} \sigma_t(\mathbf{r}(s))ds\right)$$

where t_n, t_f denote the near/far point for the ray integral; $\sigma_t(\mathbf{x}) : \mathbb{R}^3 \rightarrow \mathbb{R}$ is a neural network that models surface

density at 3D point $\mathbf{x}$; $L(\mathbf{x}, \mathbf{d}) : \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}$ is a neural network that parametrizes radiance color at 3D point $\mathbf{x}$ when observed from direction $\mathbf{d}$. In practice, the integrals above are approximated via quadrature, yielding:

$$
\begin{aligned}
C_{rf}(\mathbf{r}) \approx & \sum_{i=1}^{N} w^{(i)} L(\mathbf{r}(t^{(i)}), \mathbf{d}) \qquad (2) \\
\text{s.t} \quad & \mathbf{r}(t) = \mathbf{o} + t\mathbf{d} \\
& w^{(i)} = T^{(i)} \left(1 - \exp(-\sigma_t(\mathbf{r}(t^{(i)}))\delta^{(i)}\right) \\
& T^{(i)} = \exp\left(-\sum_{j<i} \sigma_t(\mathbf{r}(t^{(j)}))\delta^{(j)}\right) \\
& \delta^{(i)} = t^{(i+1)} - t^{(i)},
\end{aligned}
$$

where $\{t^{(1)}, \cdots, t^{(N)}\}$ are a set of sampled points on the ray that are obtained through importance sampling and $\delta^{(i)}$ is the length of the i^{th} sampling interval.

MANO and canonical representation. The MANO [37] hand model is parametrized by $\mathbf{\Theta} = \{\theta, \beta, \mathbf{p}\}$, where $\theta \in \mathbb{R}^{45}$ denotes hand skeletal pose (joint angles), $\beta \in \mathbb{R}^{10}$ hand shape (parameterized by PCA coefficients) and $\mathbf{p} \in \mathbb{R}^6$ global transformation. The MANO model then maps $\mathbf{\Theta}$ to a posed 3D mesh $\mathcal{M}(\mathbf{\Theta}) \in \mathbb{R}^{778\times 3}$. PALM-Net leverages MANO to perform inverse LBS using SNARF [5] that map points in 3D to a common canonical representation, i.e., given a 3D point in *deformed* coordinates $\mathbf{x}_d$, hand parameters $\mathbf{\Theta}$, we find the corresponding 3D point $\mathbf{x}_c$ in *canonical* space as,

$$
\mathbf{x}_d = \arg\min_{\mathbf{x}} \left|\left| \sum_{i=1}^{n_b} w_i(\mathbf{x}) \cdot B_i \cdot \mathbf{x} - \mathbf{x}_d \right|\right|_2^2, \qquad (3)
$$

where, $w_i(\mathbf{x})$ is the skinning weight associated with point $\mathbf{x}$ for bone i, and B_i represents the i^{th} bone transformation.

4.2. PALM-Net

Physically based representations: Inspired by [6, 46], PALM-Net (Figure 4c) decomposes the hand representation into a shape network $f_g(\cdot)$, a radiance field network $f_{rf}(\cdot)$, and a material network $f_m(\cdot)$. We model the canonical hand geometry with an implicit function:

$$
f_g : \mathbf{x}_c, \theta, \beta, \phi \mapsto \sigma_t(\mathbf{x}_c), z(\mathbf{x}_c), \qquad (4)
$$

where $\mathbf{x}_c \in \mathbb{R}^3$ denotes a 3D point in the canonical space, θ, β are MANO parameter, and $\phi \in \mathbb{R}^{\text{d}_\text{s}}$ captures subject-specific geometry latent code of dimension d_s. $f_g(\cdot)$ outputs the opacity value, $\sigma_t(\mathbf{x}_c) \in \mathbb{R}$, as well as geometry features $z(\mathbf{x}_c) \in \mathbb{R}^{\text{d}_\text{s}}$ for the point $\mathbf{x}_c$. Following [48], the opacity is obtained by converting the Signed Distance Function (SDF) values via the cumulative distribution function of the scaled Laplace distribution, $\Gamma_{\alpha_1,\alpha_2}(s)$, where $\alpha_1, \alpha_2 > 0$ are optimizable parameters. For details, we refer the reader to [48].

The outgoing radiance, $L(\mathbf{x}_c, \mathbf{d})$ at canonical point x_c viewed by the direction $\mathbf{d} \in \mathbb{R}^3$ is obtained as,

$$
f_{rf} : \mathbf{x}_c, z, \text{ref}(\mathbf{d}, \mathbf{n}), \mathbf{n}, \theta, \psi \mapsto L(\mathbf{x}_c, \mathbf{d}), \qquad (5)
$$

where, $\mathbf{n}$ is the surface normal obtained analytically from the SDF field, $\text{ref}(\mathbf{d}, \mathbf{n})$ reflects the view direction $\mathbf{d}$ around the normal $\mathbf{n}$, and $\psi \in \mathbb{R}^{\text{d}_\text{s}}$ is the appearance latent code.

Lastly, the spatially varying material field, f_m is used to model the physically based rendering parameters, the albedo $\alpha \in \mathbb{R}^3$, roughness $r \in \mathbb{R}$, and metallicity $m \in \mathbb{R}$ as,

$$
f_m : \mathbf{x}_c, z, \theta, \psi \mapsto \alpha(\mathbf{x}_c), r(\mathbf{x}_c), m(\mathbf{x}_c) \qquad (6)
$$

The canonical point $\mathbf{x}_c$ is encoded with hash grid encoding for all three networks, $f_{\{g,\ rf,\ m\}}$ to model high frequency details efficiently.

Physically based rendering: For physically based rendering, we follow closely [46] and compute the radiance scattered by the volume along a certain camera ray $(\mathbf{o}, \mathbf{d})$ using the quadrature approximation as,

$$
\begin{aligned}
C_{pbr}(\mathbf{r}) \approx & \sum_{i=1}^{M} w^{(i)} BRDF\Big(\mathbf{d}, \bar{\mathbf{d}}^{(i)}, \alpha\big(\mathbf{r}(\bar{t}^{(i)})\big), r\big(\mathbf{r}(\bar{t}^{(i)})\big), \\
& m\big(\mathbf{r}(\bar{t}^{(i)})\big), \mathbf{n}\Big) \cdot L_i(\mathbf{r}(\bar{t}^{(i)}), \bar{\mathbf{d}}^{(i)}) \cdot \frac{1}{pdf(\bar{\mathbf{d}}^{(i)})} \qquad (7) \\
\text{s.t} \quad & \mathbf{r}(t) = \mathbf{o} + t\mathbf{d} \\
& w^{(i)} = T^{(i)} \left(1 - \exp(-\sigma_t(\mathbf{r}(\bar{t}^{(i)})\delta^{(i)})\right) \\
& T^{(i)} = \exp\left(-\sum_{j<i} \sigma_t(\mathbf{r}(\bar{t}^{(j)}))\delta^{(j)}\right) \\
& \delta^{(i)} = \bar{t}^{(i+1)} - \bar{t}^{(i)}
\end{aligned}
$$

where $(\bar{t}^{(1)}, \bar{t}^{(2)}, ..., \bar{t}^{(M)})$ are the importance sampling offsets from the PDF estimated by radiance field samples in Equation 2; M denotes the number of samples used to approximate the integrals along the ray; $\bar{\mathbf{d}}^{(i)}$ is the incoming light direction at sampling offset $\bar{t}^{(i)}$ sampled from the distribution, $pdf(\cdot)$ (uniform distribution over the unit sphere); $BRDF(\cdot)$ denotes the simplified version of Disney BRDF [3]. The term $L_i(\mathbf{x}, \bar{\mathbf{d}})$ is the incoming radiance towards point $\mathbf{x}$ along direction $\bar{\mathbf{d}}$ and can be computed as the weighted sum of output radiance $C_{rf}(\mathbf{x}, \bar{\mathbf{d}})$ (Equation 2) and radiance emitted from an environment map $\text{Env}(\bar{\mathbf{d}})$:

$$
\begin{aligned}
L_i(\mathbf{x}, \bar{\mathbf{d}}) = & C_{rf}(\mathbf{x}, \bar{\mathbf{d}}) \qquad (8) \\
& + \exp\left(-\int_{t'_n}^{t'_f} \sigma_t(\mathbf{x} + s\bar{\mathbf{d}})ds\right) \text{Env}(\bar{\mathbf{d}}), \qquad (9)
\end{aligned}
$$

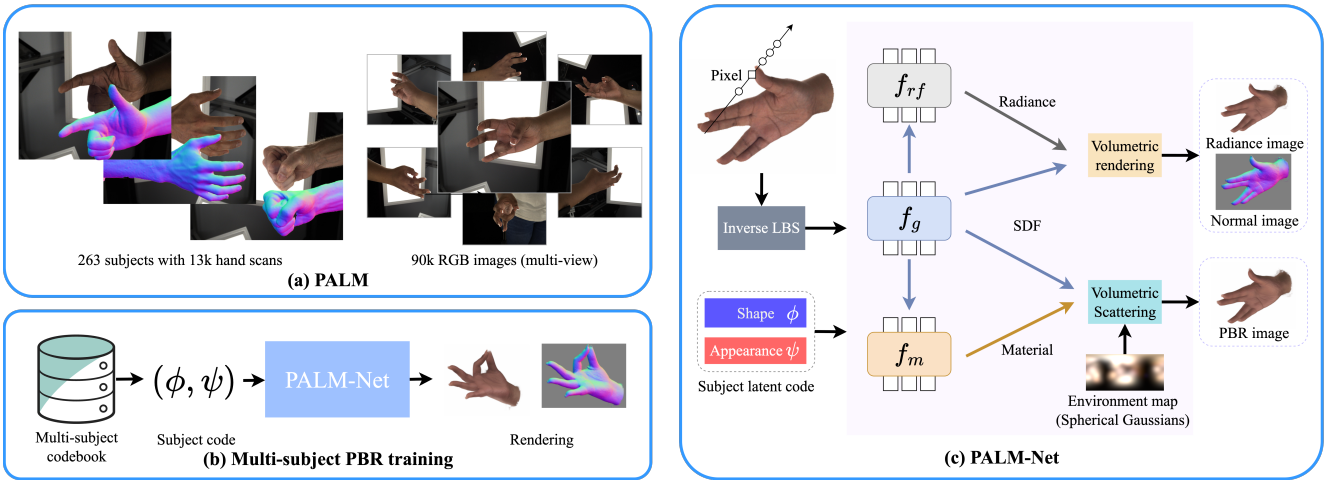


Figure 4. **PALM-Net overview**. Given (a) PALM, our multi-subject RGB dataset with 263 subjects, PALM-Net explains each subject by optimizing subject-specific shape and appearance codes (b). (c) PALM-Net is an implicit physically-based network that is conditioned on the subject codes and renders to radiance, normal, and physically-based RGB images.

where t'_n, t'_f are near and far points of integration for secondary rays. The environment map is approximated by a set of Spherical Gaussians denoted by $\mathcal{SG}_1, \mathcal{SG}_2, ..., \mathcal{SG}_G$. We refer the reader to [46] for derivation of the above equations.

Training losses: Since reconstructing geometries and material properties from RGB images is a highly under-constrained problem, we devise a loss $\mathcal{L}$ that consists of several terms. In particular, we first encourage RGB values to be consistent with an input image via

$$\mathcal{L}_{\text{rf}} = \sum_{\mathbf{r}} \left\| C_{rf}(\mathbf{r}) - \hat{C}(\mathbf{r}) \right\|, \tag{10}$$

where $\mathbf{r}$ is a ray casted from a sampled pixel on an image, and $C_{rf}(\mathbf{r})$ and $\hat{C}(\mathbf{r})$ are the rendered radiance value and ground-truth color. The PBR rendered pixels C_{pbr} are directly supervised with RGB values in a loss $\mathcal{L}_{\text{pbr}}$ similar to $\mathcal{L}_{\text{rf}}$. Since our scans provide detailed geometries, we supervise the model with the rendered scan normals:

$$\mathcal{L}_{\text{normal}} = \sum_{\mathbf{r}} \left\| \mathcal{N}(\mathbf{r}) - \hat{\mathcal{N}}(\mathbf{r}) \right\|. \tag{11}$$

Note that the geometry information is shared by PBR and radiance field. We encourage valid SDFs with the eikonal loss $\mathcal{L}_{\text{eikonal}}$ [17], which enforces the gradient at each point to have a unit norm. To encourage smooth hand surfaces, we apply a Laplacian loss $\mathcal{L}_{\text{LAP}}$ on sampled points around the hand (see SupMat). To encourage latent codes to be close to zeros, we penalize a MSE loss $\mathcal{L}_{\text{latent}}$ on both appearance and shape code. To avoid foreground model explaining background pixels, we also supervise the networks with a segmentation loss

$$\mathcal{L}_{\text{segm}} = \sum_{\mathbf{r}} \text{BCE}(\mathcal{S}(\mathbf{r}), \hat{\mathcal{S}}(\mathbf{r})) \tag{12}$$

where $\mathcal{S}(\mathbf{r}) \in \mathbb{R}$ represents the probability of a pixel being the foreground and $\text{BCE}(\cdot, \cdot)$ is the binary cross entropy loss to the ground-truth hand segmentation mask $\hat{\mathcal{S}}(\mathbf{r})$ rendered from the hand scans. Finally, to capture high-frequency details, we render image patches and compare them with the ground-truth using the perceptual similarity loss $\mathcal{L}_{\text{LPIPS}}$ [54]. The total loss $\mathcal{L}$ is defined as

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{rf}} + \lambda_{\text{pbr}}\mathcal{L}_{\text{pbr}} + \lambda_{\text{segm}}\mathcal{L}_{\text{segm}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} \\ + \lambda_{\text{eikonal}}\mathcal{L}_{\text{eikonal}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} \qquad (13) \\ + \lambda_{\text{LAP}}\mathcal{L}_{\text{LAP}} + \lambda_{\text{latent}}\mathcal{L}_{\text{latent}} \qquad (14) \end{aligned}$$

where λ_* are the weights for the loss terms (see SupMat). We gradually decrease λ_{segm} over time.

Multi-subject PBR prior: When training PALM-Net across multiple subjects, we model the detailed hand shape and appearance of each subject by conditioning PALM-Net on a shape code ϕ and an appearance code ψ. That is, we model subject identities by disentangling shapes from appearances. Empirically, we found that when training on multiple subjects, having separate latent codes for geometries and appearances yields better reconstructions than having a shared latent code for both.

Given a set of images of hands $\{\mathcal{I}\}$ from N subjects, and randomly initialized latent codes $\{\phi_{1..N}, \psi_{1..N}\}$, PALM-Net explains the images of multiple subjects by optimizing on the network weights Φ, the subject codebook $\{\phi_{1..N}, \psi_{1..N}\}$, and Spherical Gaussian parameters for the environment lights. Note that we optimize for a single environment across all subjects as the subjects are captured in the same setup. In particular, our objective function is,

$$\min_{\Phi, \{\phi_i\}_{i=1}^{N}, \{\psi_i\}_{i=1}^{N}, \{\mathcal{SG}_i\}_{i=1}^{G}} \mathcal{L} \tag{15}$$

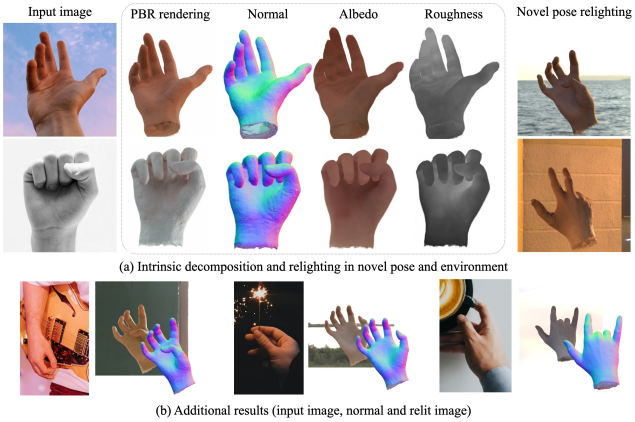


Figure 5. **In-the-wild image personalization**. (a) The first column shows the images used for personalization, followed by the renderings of the geometry and materials of the hand avatar obtained using our prior model. The PBR rendering refers to the physically-based rendering with estimated environment map. The last column shows the relighting results of personalized hand avatar in a novel pose. Our method retrieves realistic hand avatars even when the input personalization image has complex lighting effects. (b) Additional relighting results with in-the-wild images.

Personalization: A strong prior model on the hand appearance and geometry allows us to personalize our model to images of hands captured in extreme environment settings. This is mainly because the prior model constrains the albedo and material properties of the hand during personalization and all the environment effects could be explained separately. This is achieved by solving an optimization problem where the shape code ϕ and the appearance code ψ for a given input image are optimized along with the environment map while keeping the network weights Φ of the PBR prior model frozen. In particular, at each iteration, we sample a random batch of rays from the input image and optimize for the following objective:

$$\min_{\phi,\psi,\{\mathcal{SG}\}_{i=1}^{G}} \mathcal{L}_{\text{rf}} + \lambda_{\text{pbr}}\mathcal{L}_{\text{pbr}} + \lambda_{\text{segm}}\mathcal{L}_{\text{segm}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} \tag{16}$$

5. Experiments

We evaluate our baseline on the task of hand avatar personalization and relighting from a single RGB image using three different datasets. Metric and baseline details are in SupMat.

5.1. Datasets

InterHand2.6M: To evaluate the performance on real images, we use images from InterHand2.6M [30] for personalization. The dataset consists of accurate 3D hand poses of subjects in predefined poses as well as high-resolution RGB images. We evaluate on right-hand only sequences using the test split of the dataset. To reliably measure performance, we randomly select two views for each sequence, resulting in 12 sequences. We uniformly sample 20 images for each sequence for the evaluation and use the first frame of each sequence for training a personalized model.

HARP relit: Since there is no real dataset to evaluate hand avatar under novel environment and poses, following [22], we render a synthetic dataset using Blender. In particular, to create synthetic hand template, for each sequence, we sample new MANO shape parameters and apply a new skin tone using UV textures from DART [15]. To animate the hand, we use the hand pose parameters from HARP data release [22]. We use the first frame of each sequence for training a personalized model and use the remaining frames for evaluation. We render the training and evaluation images using different environment maps to evaluate relighting.

In-the-wild images: To show the generalization of our method under real world scenarios, we select in-the-wild images from the internet with diverse lighting conditions and poses. After personalization on these images, we render them with novel environment maps using novel poses from [22] and show qualitative results.

Method	PSNR↑	SSIM↑	LPIPS↓
Handy [35]	7.50	0.69	0.24
HARP [22]	9.89	0.78	0.16
UHM [32]	10.08	0.76	0.19
Ours	**12.01**	**0.84**	**0.15**

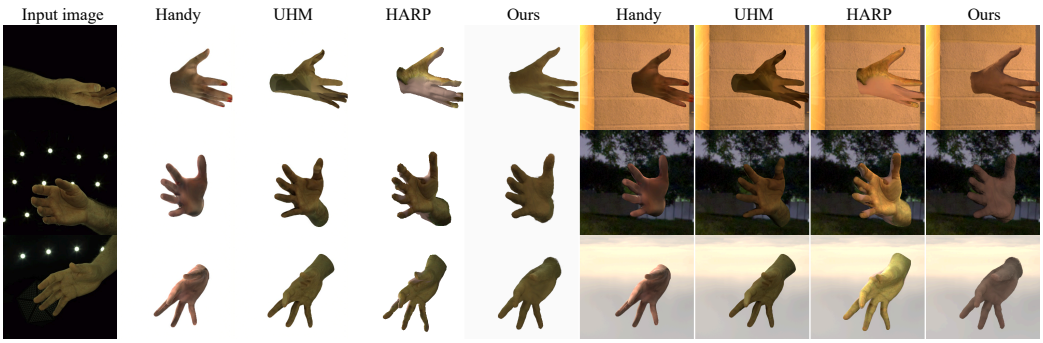
Table 2. **InterHand2.6M dataset evaluation**. Comparison of methods on single-image personalization task using PSNR, SSIM, and LPIPS metrics. Our method outperforms previous methods in the novel pose setting where the training and evaluation environment maps are the same.

Method	PSNR↑	SSIM↑	LPIPS↓
Handy [35]	12.48	0.76	0.32
HARP [22]	11.93	0.69	0.37
UHM [32]	12.30	0.74	**0.31**
Ours	**13.39**	**0.78**	0.35

Table 3. **Synthetic dataset evaluation**. Comparison of methods based on PSNR, SSIM, and LPIPS. Our method outperforms previous methods in the novel environment, novel pose setting showing that the appearance reconstructions of our model is more accurate.

5.2. Comparison and Analysis

Baseline comparison: Table 2 compares the performance of our method with the baselines on InterHand2.6M. Our method outperforms baseline methods on all metrics showing high quality rendering in novel poses and viewpoints. Figure 6 shows the qualitative images on InterHand2.6M dataset for both the training and novel environments. The images from PALM-Net are more realistic than that of the



(a) Novel poses in training environment (b) Novel poses in novel environment

Figure 6. **Personalization results on InterHand2.6M**. The first column shows the image used for personalization. (a) Personalized hand avatars rendered in novel poses in the training environment. (b) Personalized hand avatars rendered in novel poses in the novel environment. The hand avatars from our method are more realistic than other baselines.

	PSNR	SSIM	LPIPS
w/o normal	11.97	0.84	0.18
w/ normal	**12.01**	**0.84**	**0.15**

Table 4. **Effects of 3dMD normals**.

	PSNR	SSIM	LPIPS
w/o env	16.14	0.79	0.229
w/ env	**16.74**	**0.81**	**0.222**

Table 5. **Effects of modelling environment lightings**.

Figure 7. **Effects of 3dMD scans for hand personalization.**

Figure 8. **The effect of modelling environment**.

baselines. Table 3 compares ours with the baselines on the synthetic dataset, where the evaluations are performed in novel environment settings. Our PALM-Net prior model outperforms the baselines by showing more realistic relighting results in novel environments (more results in SupMat). Figure 5 shows qualitative results of personalization on in-the-wild images with diverse lighting condition and poses. Despite these challenges, our method produces realistic avatars and plausible relit images. For example, even in extreme setting where the input image is grayscale, our method can still recover plausible albedo thanks to our prior on hand appearance and the optimization over the environment map.

Supervision with 3dMD normals: Traditional multi-view techniques require a massive amount of RGB views for high-fidelity 3D reconstruction [28]. In our capture setup, we use a hybrid approach, combining 3dMD scans and sparse multi-view RGB images for training our prior models. Figure 7 shows a qualitative comparison with and without 3dMD normals. In particular, we train two multi-subject PBR prior models on PALM, one with 3dMD normal supervision and one without. Then we take these two prior models and personalize on an InterHand2.6M image. Figure 7 shows that, 3dMD normals from hand scans are crucial in reducing pepper-like artifacts and it helps to reduce floaters. Table 4 quantitatively compares in this novel pose evaluation.

Environment map optimization: During personalization, PALM-Net explains the image with geometry, material properties and environment lightings. Table 5 shows that modelling environment lightings enables the model to be more expressive in fitting the input image. An example of physically based rendered RGB images are show in Figure 8. When allowed optimizing the environment in PALM-Net, the fitting results are closer to that of the input image.

6. Conclusion

PALM is a large-scale dataset combining accurate 3D hand geometry, high-resolution multi-view imagery, and a diverse subject pool, addressing key limitations in existing datasets. Through PALM-Net, we demonstrate that physically based inverse rendering with a multi-subject prior enables realistic, relightable single-image personalization. The dataset's scale, diversity, and accompanying baseline make it a solid resource for future work.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. HOT3D: Hand and object tracking in 3d from egocentric multi-view videos. In *CVPR*, pages 7061–7071, 2025. 2

[2] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *ECCV*, pages 361–378, 2020. 2

[3] Brent Burley. Physically-based shading at disney. In *Proc. of SIGGRAPH*, 2012. 5

[4] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, pages 9044–9053, 2021. 2

[5] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, 2021. 5

[6] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gDNA: Towards generative detailed neural avatars. In *CVPR*, pages 20427–20437, 2022. 5

[7] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *CVPR*, 2023. 2, 3

[8] Zhaoxi Chen, Gyeongsik Moon, Kaiwen Guo, Chen Cao, Stanislav Pidhorskyi, Tomas Simon, Rohan Joshi, Yuan Dong, Yichen Xu, Bernardo Pires, He Wen, Lucas Evans, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, Shoou-I Yu, Javier Romero, Michael Zollhöfer, Yaser Sheikh, Ziwei Liu, and Shunsuke Saito. URhand: Universal relightable hands. In *CVPR*, 2024. 3

[9] 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*, pages 20533–20543, 2022. 3

[10] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. HMP: Hand motion priors for pose and shape estimation from video. 2024. 2

[11] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J Black, and Otmar Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. pages 1–10. IEEE, 2021. 1, 2

[12] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[13] Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, et al. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *ECCV*, pages 428–448. Springer, 2024. 1

[14] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *CVPR*, pages 494–504, 2024. 1, 3

[15] 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. In *NeurIPS*, 2022. 2, 3, 7

[16] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *CVPR*, pages 409–419, 2018. 2

[17] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. 2020. 6

[18] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, pages 3193–3203, 2020. 2

[19] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 1, 2

[20] Zhisheng Huang, Yujin Chen, Di Kang, Jinlu Zhang, and Zhigang Tu. Phrit: Parametric hand representation with implicit template. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14928–14938, 2023. 2

[21] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. pages 11–21, 2021. 2

[22] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. HARP: Personalized hand reconstruction from a monocular rgb video. In *CVPR*, 2023. 2, 3, 7

[23] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *ICCV*, pages 10138–10148, 2021. 2

[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 Trans. Graph.*, 41(4), 2022. 3

[25] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, pages 7708–7717, 2019. 4

[26] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *CVPR*, pages 21013–21022, 2022. 2

[27] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay,

Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS*, 2024. 1, 2

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 4, 8

[29] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. DeepHandMesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, pages 440–455, 2020. 2

[30] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, pages 548–564, 2020. 1, 2, 3, 4, 7

[31] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A dataset of relighted 3d interacting hands. *arXiv preprint arXiv:2310.17768*, 2023. 2, 3

[32] Gyeongsik Moon, Weipeng Xu, Rohan Joshi, Chenglei Wu, and Takaaki Shiratori. Authentic hand avatar from a phone scan via universal hand model. In *CVPR*, pages 2029–2038, 2024. 2, 3, 7

[33] 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*, pages 1154–1163, 2017. 3

[34] Akshay Mundra, Jiayi Wang, Marc Habermann, Christian Theobalt, Mohamed Elgharib, et al. Livehand: Real-time and photorealistic neural hand rendering. In *ICCV*, pages 18035–18045, 2023. 2, 3

[35] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *CVPR*, pages 4670–4680, 2023. 2, 3, 7

[36] 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*, pages 54–71. Springer, 2020. 3

[37] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6):245:1–245:17, 2017. 2, 3, 4, 5

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[39] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, pages 21064–21074, 2022. 2

[40] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, pages 4645–4653, 2017. 3

[41] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3D hand pose estimation via biomechanical constraints. In *ECCV*, pages 211–228, 2020. 2

[42] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *CVPR*, pages 3786–3793, 2014. 3

[43] three dMD Hand. 3dMDhand system series. https://3dmd.com/products/. 2, 3, 4

[44] Tze Ho Elden Tse, Franziska Mueller, Zhengyang Shen, Danhang Tang, Thabo Beeler, Mingsong Dou, Yinda Zhang, Sasa Petrovic, Hyung Jin Chang, Jonathan Taylor, et al. Spectral graphormer: Spectral graph-based transformer for egocentric two-hand reconstruction using multi-view color images. In *ICCV*, pages 14666–14677, 2023. 1

[45] 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*, 118(2):172–193, 2016. 2, 3

[46] Shaofei Wang, Bozidar Antic, Andreas Geiger, and Siyu Tang. IntrinsicAvatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. In *CVPR*, pages 1877–1888, 2024. 5, 6

[47] Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. MagicHOI: Leveraging 3d priors for accurate hand-object reconstruction from short monocular video clips. In *ICCV*, pages 5957–5968, 2025. 1

[48] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 5

[49] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. DynHaMR: Recovering 4d interacting hand motion from a dynamic camera. In *CVPR*, 2025. 1

[50] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *CVPR*, pages 4866–4874, 2017. 2, 3

[51] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. GraspXL: Generating grasping motions for diverse objects at scale. In *ECCV*, pages 386–403. Springer, 2024. 2

[52] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. pages 235–246. IEEE, 2024. 1

[53] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016. 2, 3

[54] 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. 6

[55] Xiaozheng Zheng, Chao Wen, Zhuo Su, Zeran Xu, Zhaohu Li, Yang Zhao, and Zhou Xue. OHTA: One-shot hand avatar via data-driven implicit priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 799–810, 2024. 3

[56] Andrea Ziani, Zicong Fan, Muhammed Kocabas, Sammy Christen, and Otmar Hilliges. TempCLR: Reconstructing hands via time-coherent contrastive learning. pages 627–636, 2022. 2

[57] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, pages 4913–4921, 2017. 2

[58] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, pages 813–822, 2019. 3

[59] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, 2019. 2