2025 19th International Conference on Automatic Face and Gesture Recognition (FG)

# HandOcc: NeRF-based Hand Rendering with Occupancy Networks

Maksym Ivashechkin, Oscar Mendez, Richard Bowden
CVSSP, University of Surrey, Guildford, United Kingdom
{m.ivashechkin, o.mendez, r.bowden}@surrey.ac.uk

***Abstract*— We propose HandOcc, a novel framework for hand rendering based upon occupancy. Popular rendering methods such as NeRF are often combined with parametric meshes to provide deformable hand models. However, in doing so, such approaches present a trade-off between the fidelity of the mesh and the complexity and dimensionality of the parametric model. The simplicity of parametric mesh structures is appealing, but the underlying issue is that it binds methods to mesh initialization, making it unable to generalize to objects where a parametric model does not exist. It also means that estimation is tied to mesh resolution and the accuracy of mesh fitting. This paper presents a pipeline for meshless 3D rendering, which we apply to the hands. By providing only a 3D skeleton, the desired appearance is extracted via a convolutional model. We do this by exploiting a NeRF renderer conditioned upon an occupancy-based representation. The approach uses the hand occupancy to resolve hand-to-hand interactions further improving results, allowing fast rendering, and excellent hand appearance transfer. On the benchmark INTERHAND2.6M dataset, we achieved *state-of-the-art* results.**

## I. INTRODUCTION

As one of the most expressive parts of the human body, the hands play a crucial role in communication, interaction, and manipulation tasks, which drives the necessity for accurate and versatile hand estimation. Hand pose estimation, as well as hand synthesis and rendering, is important to many areas, including: human-computer interaction, avatar generation, sign language production, and augmented or virtual reality applications such as teleoperation or telepresence.

There has been a significant body of work devoted to 3D hand pose estimation over the years. The most prominent works are often monocular, exploiting image convolution, *e.g.,* [32, 42, 43, 48]. However, leveraging advances in technology, particularly GPU acceleration, enables us to achieve volumetric hand reconstruction alongside 3D rendering for novel view synthesis.

Most existing methods employ sparse 3D skeletal hand estimation, and for hand rendering they exploit mesh-based parametric representations such as MANO [41]. MANO parameterizes a 3D hand with a set of angles and shape coefficients that incorporate forward kinematics and generate a realistic hand mesh. It allows for the efficient estimation of the 3D volume of the hand, and the corresponding hand mesh can be used for rendering. These properties, alongside a differentiable implementation, have made MANO a popular and widely used hand model.

Traditional hand-rendering methods often rely on texture maps and a colored mesh, where the hand geometry is controlled by a kinematic model, examples of such approaches were demonstrated in [10, 12, 50]. Nevertheless, such methods have drawbacks. For instance, relying on meshes can lead to mesh artifacts, limits on fidelity/detail, costly generation of personalized texture maps, and challenges in handling self-occlusions and intersections.

Recently, the neural radiance field (NeRF) [31] has gained a lot interest due to its ability to represent the volume density and color space as a continuous function. Despite their recent popularity and speed, NeRFs are still attractive as they demonstrate excellent generalization to novel-view synthesis. This is due to their use of a continuous function. While the original proposal for NeRF was for static scenes, a lot of work has since explored extensions capable of integrating dynamics into the NeRF model.

### A. Related Work

The articulation of the human body, especially hands, presents numerous challenges for neural rendering, particularly when generalizing across complex shapes and motions. Some of the first approaches to adapting NeRF to a dynamic scene were D-NeRF [40] by Pumarola *et al*. and Nerfies [35] by Park *et al*. Both methods are similar in the way they introduce a deformation field (*e.g.,* MLP network) that learns the transformation from a target scene to canonical space, and a canonical NeRF model that predicts colors and density. The authors argue that a two-model approach that introduces a transformation from observation to canonical space is better than direct estimation and helps in generalization. However, this technique also imposes additional constraints on models to learn information about the shared geometry between the canonical and observation space, and the corresponding appearance.

To achieve better reconstruction and thus rendering for a human body, parametric models have gained popularity. They allow integration of human geometry (*e.g.,* kinematics) into a neural model and can achieve more accurate and faster optimization. We can divide them into two main types of human body parameterization: implicit or mesh-based representations.

The vital advantage of an *implicit* parameterization is that it can be represented with a continuous and differentiable function. Examples are the signed distance field (SDF) used in [1, 2, 20, 34, 45], occupancy maps [30], implicit surfaces, point clouds, and transformation fields [46]. For faster convergence and better accuracy, the implicit models are normally conditioned with input data such as sparse points (*e.g.,* skeletons) or mesh parameters. The imGHUM approach [1] of Alldieck *et al*. utilizes skeleton points from

979-8-3315-5341-8/25/$31.00 ©2025 IEEE

the human body to condition the SDF. Similarly, Karunratanakul *et al*. in HALO [19] exploit the hand skeleton to condition an occupancy network. The NASA model [9] by Deng *et al*. presents a neural pose-conditioned occupancy approach based on a mesh, where high-quality surface details are learned using per-bone deformable transformation. The PTF [46] of Wang *et al*. extends NASA's approach to learning occupancy functions in the continuous rest-pose space by exploiting piece-wise transformation fields. Although NASA uses SMPL [28] mesh parameters as initialization, the PTF tries to robustly fit a mesh to the predicted point cloud.

*Mesh* representations for the human body are extremely popular due to their differentiability, simplicity, and compactness. They have been proposed for body (SMPL), hands (MANO [41]), faces (FLAME [27]), and even animals [54]. One of the main advantages of a mesh-based model is that it provides a volumetric shape that exhibits aspects of realism. The SMPL model is often utilized in body capture from images (*e.g.,* [25, 36]), and MANO in various hand estimation problems (*e.g.,* [4, 53]). However, the main problems with meshes are their coarse structure and low resolution, which are especially evident when it comes to mesh rendering. One way to mitigate these issues involves increasing the number of faces to provide high-fidelity meshes [7, 29], at the cost of increasing computational complexity. PHRIT [14] combines the advantages of a MANO mesh and implicit SDF to obtain a high-fidelity reconstruction at infinite resolution but lacks real-time inference. Other techniques addressed the UV texture map of the mesh to improve hand color/texture [6], or exploit graph convolution neural networks to obtain richer information about the hand surface [10].

Implicit models tend to take longer to converge, and most of the current implementations try to combine both (implicit and mesh) representations to exploit their advantages. Moreover, with the continuous property of the NeRF that enables learning of density and color, hand-rendering methods provide accurate results with real-time efficiency. One of the first approaches to integrate NeRF for 3D hand rendering was LISA [8] of Corona *et al*. LISA exploited MANO parameters together with local bone coordinates to predict per-bone signed distance and color. The signed distances contribute to the final volume densities that allow rendering. However, due to the complexity of the mesh, the method struggles to perform in real-time.

HandNerf [11] by Guo *et al*. presents a framework for 3D hand rendering utilizing NeRF. It uses a deformation field to transform the input scene into a canonical space. Additionally, HandNerf exploits a MANO mesh to query the closest facet for a 3D point, which is used to predict texture colors. Similarly, LiveHand [33] by Mudra *et al*. uses mesh textures along with the distance to the mesh surface as an input to a NeRF to estimate hand density and color. Both LiveHand and HandNerf exploit rendered MANO mesh depth to provide an extra loss and awareness to the NeRF model, along with the ray bounds determined by a 3D hand mesh volume. The Hand Avatar [7] of Chen *et al*. provides a high-resolution MANO-HD mesh with more faces and vertices. Their method proposes a shading field, where anchors are used on the mesh to extract albedo information of the hand poses. Similarly, HARP [18] explicitly model a parametric mesh-based hand with a normal map and albedo to tackle lightning conditions and articulation.

The hand appearance was mainly tackled by an implicit function that learns hand texture from multi-view images [8, 11]. LiveHand embeds hand textures on the MANO UV texture map, while HandAvatar and HARP exploit the albedo and normal map from a MANO mesh. Handy [38] utilizes a GAN model to generate high-fidelity UV hand mesh textures. In contrast, our work employs a separate Convolutional Variational Autoencoder (CVAE) to explicitly extract latent hand texture features directly from the desired image. These extracted features are then used to condition the rendering model, allowing for more control and flexibility.

### B. Motivation and Novelties

We propose a framework for novel pose and novel view high-fidelity hand rendering. Despite recent advancements in NeRF and rendering via Gaussian splatting, hand rendering remains an unsolved problem due to numerous challenges (*e.g.,* high motion, finger interactions, etc.). The current *state-of-the-art* approaches rely heavily on the MANO mesh model, which introduces significant limitations: if the mesh is poor, the rendering quality suffers as well. Additionally, obtaining accurate hand meshes is inherently challenging and often imprecise, requiring rendering and fitting to multi-view data. These meshes are restricted by low resolution and coarse surface detail, which can lead to rendering artifacts.

To avoid reliance on a parametric mesh model like MANO, we instead leverage an implicit shape representation by probabilistically modeling the occupancy of the hand. This paper also serves as a proof of concept, demonstrating that an implicit model without mesh information achieves *state-of-the-art* performance. To the best of our knowledge, we are the first to render a dynamic hand using a pose-conditioned NeRF without relying on an underlying MANO model. The main input to our model is as simple as a sparse 3D hand skeleton, which is easier to obtain than a volumetric mesh (using a basic triangulation).

Furthermore, we remove the fundamental necessity for a parametric structure, which limits the application of current approaches to objects or body parts for which such models are unavailable. Many existing methods rely heavily on explicit mesh textures, thus restricting their applicability to areas where models exist. By adopting an implicit shape representation, we avoid these constraints and open oportunties for more flexible, extendable, and accurate modeling.

We evaluate our model on the benchmark INTERHAND2.6M dataset [32], achieving *state-of-the-art* results. Additionally, by bypassing meshes, we propose a number of novelties and advantages that include:

1) Utilizing an occupancy map to facilitate efficient hierarchical sampling of a hand surface for NeRF rays.
2) Conditioning the NeRF with CNN embeddings to provide improved hand appearance and shape transfer.

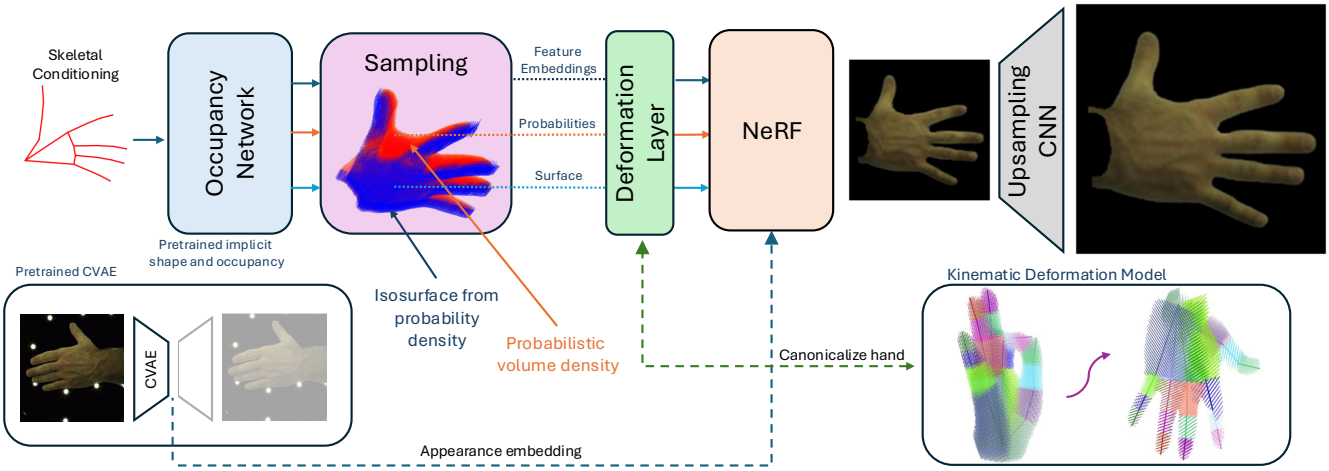



Fig. 1: This figure demonstrates an overview of the proposed approach. Samples are drawn from the occupancy network which is conditioned on the skeletal input. The occupancy model returns per-point probabilities and features of the surface points. The surface points go through a deformation layer that canonicalizes the hand input. Afterwards, per-point occupancy encodings are then given to the NeRF (MLP) along with appearance embeddings. The NeRF renders an RGB image with corresponding features. The final CNN layer upsamples and refines the NeRF output.

3) Tackling interacting hands by modeling occupancies that prevent hand intersection.
4) Efficient and fast rendering, where the hand geometry implicit to occupancy map is learned via a NeRF on a downscaled image, while exploiting a CNN for accurate image upsampling.

## II. METHODOLOGY

The overview of the proposed pipeline is demonstrated in Fig. 1. The ultimate goal is an accurate 3D rendering of an articulated hand from a single view. We achieve this using a pre-trained occupancy network, a sparse 3D skeleton, and a NeRF renderer. Additionally, we exploit appearance embeddings extracted with an image model, and an upsampling CNN that improves NeRF output.

In the following sections, we discuss each part of the pipeline in turn. First, we cover the point cloud extraction and representation of the hand occupancy model. Next, we present hand appearance and shape transfer using a CVAE model. After that, we outline volumetric hand rendering with NeRF and the efficient hierarchical sampling of the hand surface. We also discuss the importance of the deformation model and canonical representation. Finally, we describe an upsampling CNN model that increases the rendered image quality.

### A. Point Cloud Extraction

The first step in our pipeline is to extract dense point clouds from multi-view hand data to train an occupancy network. We assume $N$ camera views with known calibration parameters $\mathcal{P}$. Given an approximate 3D hand bounding box, we generate a uniformly distributed point cloud and use projection matrices $\mathcal{P}$ to find the corresponding 2D projections of the 3D points.

Assuming color consistency of the hand pixels among images, we filter the points to identify those that correspond to the hand. To support this assumption, we normalize images for luminance, brightness, color, contrast, etc. This normalization ensures minimal color discrepancy across authentic hand projections. We can test this by calculating the standard deviation over $N$ views for each point, with low variance indicating true hand projections.

While this filters hand points, background points with consistent colors may remain. We re-project all 3D points back into the 2D images and check their location with 2D binary hand masks to further filter incorrect points.

### B. Hand Occupancy Representation

Given the dense point clouds for each hand, the next step is to train an occupancy network. The aim is to learn an approximate hand shape by conditioning the occupancy network with a sparse hand skeleton.

The occupancy network provides an implicit hand representation, where for each 3D point, it returns the probability of whether a point belongs to the hand. If $\mathbf{x} \in \mathbb{R}^3$ is an arbitrary 3D point, $\mathbf{S} \in \mathbb{R}^{n\times 3}$ a sparse hand skeleton of $n$ points, then occupancy network $\mathbf{O}$ can be viewed as a probability function such that $\mathbf{O} : \mathbb{R}^3 \times \mathbb{R}^{n\times 3} \rightarrow [0, 1]$.

We canonicalize the occupancy model to align with the same 3D space centered at the origin. Let $\mathbf{s}_0$ be the first point in the matrix $\mathbf{S}$ and the root joint of a hand (*e.g.,* wrist). The normalized hand skeleton $\tilde{\mathbf{S}} = \mathbf{S} - \mathbf{1}_n \mathbf{s}_0^T$ conditions the occupancy map. Note that $\tilde{\mathbf{s}}_0$ is at the origin, *i.e.,* $\tilde{\mathbf{s}}_0 = 0$. Moreover, we represent a left hand as a flipped right hand by mirroring points along the $x$-axis.

It is crucial to efficiently exploit the occupancy network in the case of interacting hands, especially to determine the probability of intersection. $\mathbf{S}_R$ and $\mathbf{S}_L$ denote sets of points for the right and left hand, respectively. $\mathbf{P}$ is a matrix of

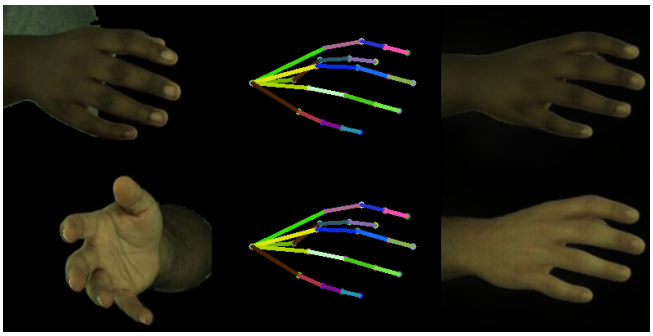


Fig. 2: This figure demonstrates the results of the CVAE model. The first column shows input images $\mathbf{I}^A_{H_i}$ and $\mathbf{I}^B_{H_j}$ of two different hands from different people. The second column shows the same hand skeleton $\mathbf{H}^X_k$ rasterized on an RGB image. The output of the model is the last column, showing synthetically generated images $\mathbf{I}^A_{H_k}$, $\mathbf{I}^B_{H_k}$. These images closely resemble the input persons identity in terms of shared features, albeit with different skeleton shape.

$K$ 3D points spanning both hands, and $\mathbf{t} = \mathbf{s}^L_0 - \mathbf{s}^R_0$ is the offset between the hands. The equation to determine the occupancy probability for the right hand is $\mathbf{O}(\mathbf{P}_i - \mathbf{s}^R_0, \tilde{\mathbf{S}}_R)$ for all indices $i \in \{1, \ldots, K\}$ of matrix $\mathbf{P}$. For the left hand, the input points $\mathbf{P}$ must not only be shifted but also flipped and translated by the hand offset to preserve the original relation of the left to right hands. Let $f$ be a function that flips input points by multiplying the $x$-axis by -1. Then the occupancy map for the left hand of the same 3D point set $\mathbf{P}$ is:

$$\mathbf{O}(f(\mathbf{P}_i - \mathbf{s}^L_0), f(\tilde{\mathbf{S}}_L)) \tag{1}$$

The probabilities can serve as labels and strong indicators of which specific hand the points belong to, potentially determining a possible hand-to-hand intersection. Crucially, the occupancy probabilities help set the boundaries of the hand and allow for the extraction of hand surfaces, which will be discussed in the following sections.

### C. Hand Appearance and Shape Transfer

Although the occupancy network is trained on point clouds derived from multi-view data, in practice, only a single image is provided as input. Therefore, the hand parameters such as shape and appearance (skin color, wrinkles, gender, hair, nails, etc.) have to be extracted from a single image.

Many approaches parameterize the shape of the hand with MANO which exploits principal component analysis (PCA) to model hand shape. As is common in the literature [23, 26, 53], mesh parameters are estimated from input images by supervising network training using ground truth mesh parameters. However, since we assume no mesh parameterization in the proposed pipeline, we demonstrate a new, alternative approach that extracts both shape and appearance from the images. Moreover, it is also suitable for hand shape and appearance transfer.

First, let $\mathbf{H} \in \mathbb{R}^{n \times 2}$ be a 2D hand skeleton found by projecting the 3D hand skeleton $\mathbf{S}$, using known projection parameters $\mathcal{P}$ and $\mathbf{I}$ is an RGB image of a hand. The objective is to learn a function $\phi$ capable of transforming a hand skeleton $\mathbf{H}$ and image $\mathbf{I}$ into a latent vector representing the hand composition. This composition should capture both the skeletal structure of the hand $\mathbf{H}$ and its appearance in the image $\mathbf{I}$.

For this task, we exploit a convolutional variational autoencoder (CVAE) [22], because they are efficient generative models capable of accurately encoding the input into a latent representation that follows a Gaussian distribution. The CVAE aims to take the hand skeleton $\mathbf{H}$ and an image of a hand $\mathbf{I}$, then compress that information into a lower dimensionality latent space with sufficient detail that a decoder can reconstruct an image of a hand with the same appearance $\mathbf{I}$ and skeleton $\mathbf{H}$.

Let the function $\phi$ be the CVAE encoder, where the function $\psi$ decodes the latent vector of $\phi$ to an RGB image. Then for any person $X$ and for any hand pair $(i, j)$:

$$\phi(\mathbf{H}^X_i, \mathbf{I}^X_{H_i}) = \phi(\mathbf{H}^X_i, \mathbf{I}^X_{H_j}) \ \wedge \ \psi\big(\phi(\mathbf{H}^X_i, \mathbf{I}^X_{H_j})\big) = \mathbf{I}^X_{H_i} \tag{2}$$

Where, $\mathbf{H}^X_i$ is $i$-th hand skeleton of the $X$-th person, and $\mathbf{I}^X_{H_j}$ is an image of the $\mathbf{H}_j$ hand of the $X$-th person. In other words, the latent space for the same hand skeleton and images of the same person are equal. For two different identities $A$ and $B$, the goal of the CVAE is to be able to produce the following:

$$\psi\big(\phi(\mathbf{H}^A_i, \mathbf{I}^B_{H_j})\big) = \mathbf{I}^B_{H_i} \tag{3}$$

Which is a hand image generated of person $B$ with a hand skeleton of person $A$. The CVAE is forced to disentangle the skeleton pose from the hand appearance of the image. Consequently, by providing different people in the training set, the CVAE can generalize over various shapes and appearances.

Since each person has a different distribution of hand skeletons, in practice, Eq. (3) cannot be used directly, because such images may not exist. Therefore, the CVAE necessitates input from the same individual, as specified in Eq. (2). The associated losses for training serve to enforce consistency within a single person's latent space. They comprise the image loss of the CVAE decoder, and the Kullback-Leibler divergence [17]. Examples are demonstrated in Fig. 2.

### D. Hand Rendering

The standard NeRF architecture processes a 5-dimensional input, consisting of 3D point coordinates, $\mathbf{x}$, paired with viewing direction, $\mathbf{d}$ (represented as a 3-dimensional vector). It then estimates volume density $\sigma(\mathbf{x})$ alongside the RGB color vector $C(\mathbf{r})$. Leveraging multiple views, NeRF casts rays $\mathbf{r} = \mathbf{o} + t\mathbf{d}$ from pixel coordinates using camera parameters $\mathcal{P}$, where $\mathbf{o}$ denotes the camera origin. Subsequently, it samples these rays within predefined bounds $t \in [t_{\min}, t_{\max}]$. The predicted colors are accumulated from the colors and volume densities along the corresponding camera rays.

$$C(\mathbf{r}) = \int_{t_{\min}}^{t_{\max}} T(t)\, \sigma\big(\mathbf{r}(t)\big)\, \mathbf{c}\big(\mathbf{r}(t), \mathbf{d}\big)\, dt \tag{4}$$

$$T(t) = \exp\left( - \int_{t_{\min}}^{t_{\max}} \sigma\big(\mathbf{r}(s)\big) ds \right) \tag{5}$$

Where, the $T(t)$ function corresponds to the accumulated transmittance along the ray, and $\mathbf{c}$ is a color function of the ray and viewing direction. In the literature, the continuous NeRF function is implemented via an MLP. In practice, the integral is approximated by weighing the discrete point samples along the ray. To mitigate the dependency on the fixed ray bounds, hierarchical volume sampling is performed based on a coarse density estimate.

We build our framework around the concept of NeRF's rendering approach, however, we adapt it to the articulated hand problem. Hand motion is governed by a skeletal structure, while appearance and shape are determined by CVAE image embeddings $\mathcal{A} \in \mathbb{R}^n$. Directly encoding a skeleton vector and 3D coordinates into a NeRF MLP is inefficient and results in poor conditioning, as the skeletal embeddings lack volumetric information, are sparse, and lead to poor generalization (*e.g.,* see also [11]). Therefore, we leverage a pre-trained occupancy decoder features $\mathcal{F} \in \mathbb{R}^m$, where the input 3D points have already been processed via a skeleton occupancy embedding. These features encapsulate spatial information and the relationship to the 3D skeleton, serving as input to the NeRF.

Furthermore, the occupancy, combined with per-point features, yields probabilities $\mathcal{P}$ that we employ as density cues to expedite NeRF convergence. This is particularly beneficial for interacting hands, where probabilities facilitate the identification of specific hands by selecting the maximum probability.

Refining the NeRF output via a CNN model is one of our objectives, and several studies (*e.g.,* [11, 33]) enforced NeRF to produce additional $d$-dimensional features $\mathbf{f}_N \in \mathbb{R}^d$ alongside RGB colors and volume density. This approach is adopted because simple RGB channels lack high-dimensional information regarding volumetric shape and details, which are crucial for the post-processing CNN model. Consequently, we incorporate this strategy into our framework as well. Ultimately, NeRF can be conceptualized as a function $\boldsymbol{g}$ such that it takes appearance, occupancy features and probabilities, and returns color, density, and volumetric features:

$$\boldsymbol{g} : (\mathcal{F}, \mathcal{A}, \mathcal{P}) \rightarrow (\mathbf{c}, \sigma, \mathbf{f}_N) \tag{6}$$

In the original NeRF formulation, the authors randomly sample rays from images. However, such a strategy restricts us to an RGB loss (*e.g.,* mean squared error, MSE) on the predicted and ground truth pixel colors. To integrate a perceptual image loss that is superior in capturing detailed features (*e.g.,* LPIPS [49]) we use all image rays to fully render an image. Nevertheless, loading all rays is very computationally expensive. Therefore, for efficiency reasons, we prune rays with zero occupancy probability. This allows us to perform batching over the images without a significant memory demand.

### *E. Hand Bounds*

In the vanilla NeRF, the ray sampling bounds are set by user-defined nearest and farthest distances that span the object. The "coarse" NeRF model uses uniformly sampled points along rays, and a "fine" model then provides hierarchical sampling.

For dynamic hands, using fixed bounds is inefficient because they must be large enough to span over all hand articulation, and to span the whole object's volume, many samples are required. Therefore, we exploit the probabilistic occupancy network to establish constraints and sample points in close proximity to the hand surface. We define the origin of the ray intersecting the hand surface based on an occupancy probability threshold $p_{\min}$ (e.g., 0.1). The upper bound of the ray can be determined by the point of maximum occupancy saturation (*i.e.*, $p_{\max}$, close to probability 1.0), a predefined minimum distance (e.g., 1-2 cm), or, as implemented in our approach, a combination of both. The estimated hand surface is essentially the closest point on a ray to the camera that has an occupancy probability equal to a pre-defined threshold. Figure 1 depicts the hand surface represented by blue points.

### *F. Occupancy Hierarchical Sampling*

The probabilities returned by the occupancy network can be interpreted as densities in a NeRF model. However, these occupancy probabilities are not reliable enough to directly replace the NeRF density field, as doing so would lead to degraded rendering quality. Instead, we leverage the occupancy field to incorporate additional sample points from regions with the highest probabilities by employing hierarchical sampling [31]. This procedure mirrors the original proposal by the NeRF authors, which involved a fusion of "coarse" and "fine" models. However, by sampling directly from the hand surface and augmenting it with points of the highest probabilities, we circumvent the need for a "fine" model.

### *G. Deformation Model*

To improve generalization and accelerate training, we integrated a deformation model that transforms an observed hand pose into a canonical one. This assists NeRF in optimizing a single canonical pose, rather than various hand poses, thereby reducing complexity. We base our deformation only on skeletons. Firstly, we determine the per-bone rotations and translations between the canonical and observed sparse skeletons. Then, we segment a hand point cloud based on proximity to the underlying skeleton edges and transform the points using corresponding per-bone rotations and translations. Unlike other methods that utilize MLP-based deformation for enhancement or prediction, our approach does not require a deformation neural network. The per-bone rigid transformations are precise, and any artifacts resulting from the segmentation are either negligible or mitigated by filtering the point cloud through the occupancy network.

### H. CNN Upsampling

As is often mentioned, NeRF can be slow to train and takes a long time to fully converge. Even with the proposed efficient sampling, the time to render a full-resolution image takes significant computational time. Since NeRFs are capable of accurately representing volumetric geometry and high rendering quality, these properties can be exploited at a much lower image resolution. Moreover, upscaling CNN models have shown excellent capability in restoring and refining down-scaled images. Therefore, we enhance rendering speed by training a NeRF on low-resolution hand geometry and subsequently reconstructing the full-size image using a CNN. This process involves providing a provisional RGB image along with additional feature channels $\mathbf{f}_N$.

## III. EXPERIMENTS

We primarily focus our evaluation on the publicly available INTERHAND2.6M benchmark dataset. It is one of the largest hand datasets containing 2.6 million images captured with 140 cameras of 26 unique people. The dataset is commonly used to evaluate 3D hand estimation and rendering.

### A. Implementation Details and Reproducibility

To extract hand point clouds from multi-view images, we combined hand masks and color consistency to filter out outliers. The images are normalized with histogram equalization and converted to the HSV format. Since ground truth point clouds are unavailable, visual verification on a random subset of hands confirms their high realism. For our occupancy network, we leverage a PointNet encoder [5] to condition it with the sparse 3D input skeleton data. The decoder utilizes conditional batch normalization with ResNet [13] blocks to convolve the input 3D points, incorporating the embedded skeleton from the encoder, and transforming the points into logits. The intersection over union accuracy of our occupancy network is 80% on the validation set. Such performance demonstrates high overlap with the ground truth hand occupancy, enabling reliable shape reconstruction.

For the CVAE model, we used a CNN with residual blocks to downsample and upsample the images. The latent space estimated by the encoder is further reduced by PCA. The loss for the CVAE combines LPIPS, L1 and KL Divergence. For the NeRF, we use 8 uniformly distributed point samples along the ray and another 8 points from hierarchical occupancy sampling. The NeRF has a width of 256-512 hidden units for the MLP, with a depth of 8 layers. The CNN upscaling model is SRResNet [24] and has a scale factor of 2. The model is trained end-to-end using a combined LPIPS (0.4) and L1 (0.6) loss on the images. All models were trained until convergence using an Adam optimizer [21] with an NVIDIA GeForce RTX 3090 GPU. At inference, our unoptimized model renders at 7fps in comparison to a simple mesh render at 58fps.

### B. Affine Image Transform

Using the original downscaled image in NeRF is inefficient, as it often contains a large amount of empty background, with the hand occupying only a small portion of the space. Using the approximate 2D bounds of the hand (found from a 2D skeleton), we crop the original image into a fixed bounding box via an affine transform $\mathbf{T} \in \mathbb{R}^{3\times3}$.

During training, we introduce a random scale or translational shift to the affine transform $\mathbf{T}$. Such an augmentation casts new rays, providing subpixel precision. It helps the NeRF to generalize and provides new observations rather than overfitting to the same rays every time. To preserve the original image coordinates, we update the corresponding intrinsic matrix. The high-resolution bounding box images are used to train the super-resolution CNN model. To restore the original image, the transform $\mathbf{T}^{-1}$ is applied to the cropped image.

TABLE I: Comparison of rendering quality on INTERHAND2.6M. The symbol * indicates that the method was the version implemented in the LiveHand paper [33].

| | PSNR ↑ | LPIPS (x1000) ↓ ↑ |
|---|---|---|
| Mesh wrapping | 28.28 | 49.44 |
| SMPLpix [39] | 32.37 | 26.57 |
| A-NeRF* [44] | 28.07 | 94.41 |
| LISA* [8] | 29.36 | 78.46 |
| LiveHand [33] | 32.04 | **25.73** |
| Ours | **32.38** | 27.92 |

### C. Quantitative evaluation

Table I compares our method against the *state-of-the-art* approaches. The evaluation protocol follows the instructions outlined in the LiveHand description [33]. We outperform other approaches except for the LPIPS metric. This is because the ground truth images are masked with the MANO mesh, and since our method does not utilize mesh information, the LPIPS metric becomes sensitive to the overall hand shape.

We followed the evaluation guidelines in HandNeRF [11] and evaluated model performance trained with 4, 7, and 10 views on 18 different test views. The results of the experiments are shown in Table II. We outperform the *state-of-the-art* except for the SSIM metric, where we are comparable.

By applying the training and evaluation protocol from the HandAvatar [7] method, the comparison results on different splits of the INTERHAND2.6Mdataset are presented in Table III. On most metrics, our approach is either better or comparable to the *state-of-the-art*.

The results on HanCo [51, 52] are shown in Table IV. We used the HandAvatar code and trained the method on all images of persons 26 and 29 without hand-to-object interactions. The dataset is very challenging and has a lot of illumination changes, hence, the perceptual accuracy is rather low. However, our method outperforms the HandAvatar method on the PSNR and is marginally behind on LPIPS and SSIM metrics.

For a fair evaluation, the main input to our rendering model is a 3D skeleton provided in the dataset, as the competitors use the ground truth MANO mesh parameters.

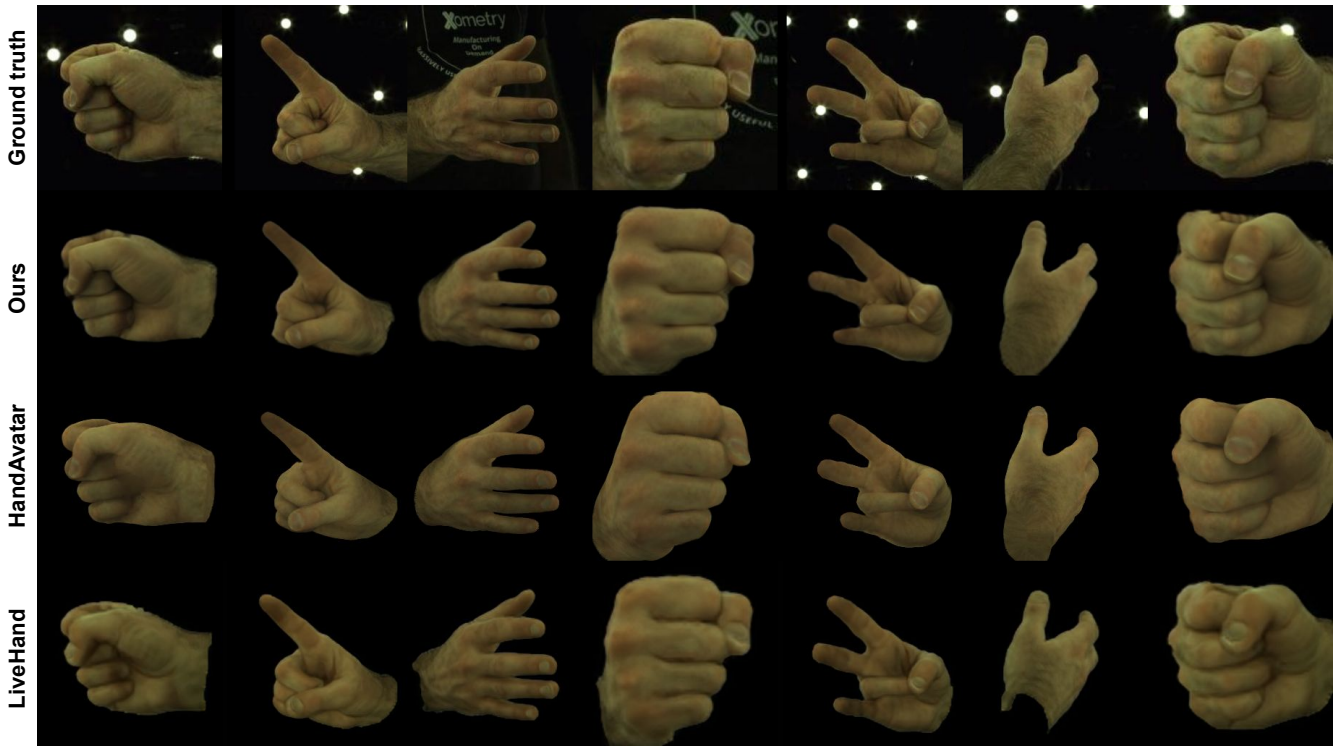



Fig. 3: Comparison of single hand rendering to LiveHnad [33] and HandAvatar [7] methods.

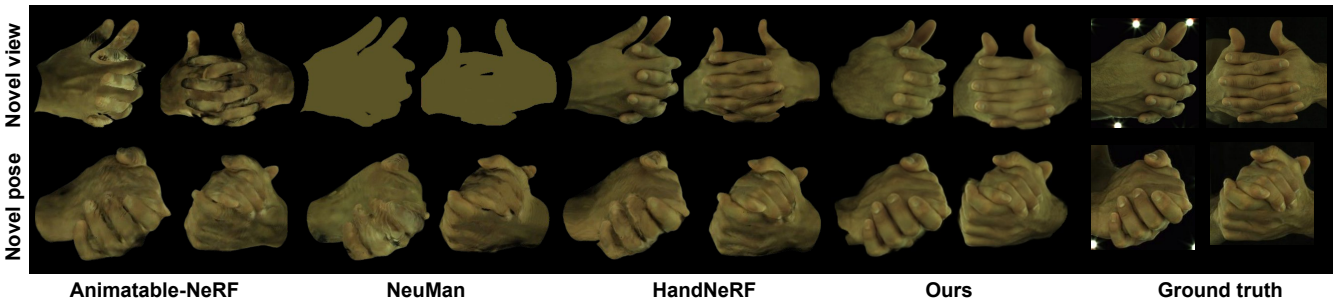
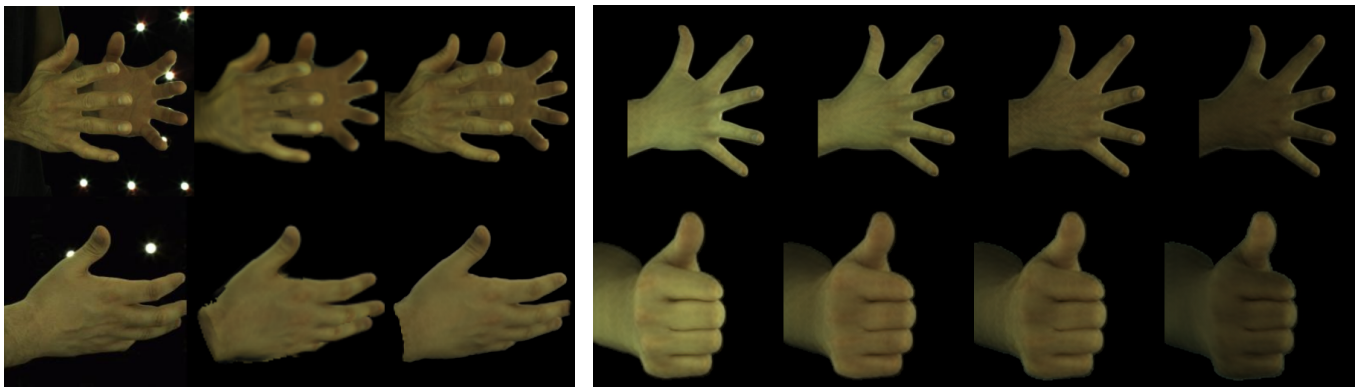



Fig. 4: Comparison of interacting hands with Animatable-NeRF [37], NeuMan [16], and HandNeRF [11]. Each pair of images shows two different views of the same pose. The first row applies NeRF to a novel view, while the second row applies it to a novel pose. The competitor images are sourced from the HandNeRF.



(a) (b)

Fig. 5: Figure (a) shows: ground truth, output from the NeRF, output from the upscaling CNN. Figure (b) demonstrates appearance transfer of the same hand pose to different identities.

TABLE II: Comparison of rendering results on single and interacting hands. The columns represent models trained on 4, 7, and 10 views, respectively, and tested on 18 views. The symbol * denotes models implemented by HandNeRF [11]. The symbol × indicates that the model failed to converge.

| | 4 views | | | 7 views | | | 10 views | | |
|---|---|---|---|---|---|---|---|---|---|
| | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS |
| Single hand | | | | | | | | | |
| Pose-NeRF* [3] | 27.085 | 0.935 | 0.092 | 29.264 | 0.930 | 0.070 | 29.212 | 0.939 | 0.073 |
| Ani-NeRF* [37] | 30.260 | 0.958 | 0.070 | 31.642 | 0.963 | 0.058 | 31.778 | 0.968 | 0.062 |
| NeuMan* [16] | 30.342 | 0.959 | 0.069 | 31.236 | 0.962 | 0.057 | 31.841 | 0.970 | 0.055 |
| HandNeRF [11] | 31.049 | **0.965** | 0.058 | 31.855 | **0.969** | 0.045 | 32.703 | **0.974** | **0.037** |
| Ours | **31.958** | 0.962 | **0.043** | **32.741** | 0.965 | **0.039** | **33.090** | 0.967 | 0.038 |
| Interacting hands | | | | | | | | | |
| Pose-NeRF* [3] | 25.019 | 0.874 | 0.187 | 27.241 | 0.901 | 0.138 | 27.646 | 0.916 | 0.107 |
| Ani-NeRF* [37] | 28.032 | 0.941 | 0.086 | 28.854 | 0.944 | 0.084 | 29.357 | 0.949 | 0.079 |
| NeuMan* [16] | × | × | × | × | × | × | × | × | × |
| HandNeRF [11] | 29.035 | **0.955** | 0.084 | 30.069 | **0.962** | 0.081 | 30.757 | **0.956** | 0.072 |
| Ours | **29.746** | 0.931 | **0.079** | **30.706** | 0.938 | **0.071** | **30.836** | 0.939 | **0.070** |

TABLE III: Comparison of rendering quality on the HandAvatar [7] splits (first row) of the INTERHAND2.6M dataset. A higher SSIM is better.

| Method | *test/Capture0* | | | *test/Capture1* | | | *val/Capture0* | | |
|---|---|---|---|---|---|---|---|---|---|
| | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM |
| SelfRecon [15] | 0.142 | 26.38 | 0.878 | 0.138 | 25.18 | 0.875 | 0.149 | 25.78 | 0.868 |
| HumanNeRF [47] | 0.114 | 27.64 | 0.883 | 0.117 | 26.31 | 0.880 | 0.119 | 27.80 | 0.881 |
| HandAvatar [7] | **0.103** | 28.23 | 0.894 | **0.107** | 26.56 | **0.890** | **0.106** | 28.04 | 0.890 |
| Ours | 0.107 | **28.33** | **0.895** | 0.117 | **26.62** | 0.886 | 0.112 | **28.23** | **0.891** |

TABLE IV: The comparison results on the HanCo [51, 52] dataset.

| | PSNR ↑ | LPIPS (x1000) ↓ | SSIM ↑ |
|---|---|---|---|
| HandAvatar [7] | 18.712 | **16.231** | **0.844** |
| Ours | **19.181** | 16.320 | 0.842 |

The skeleton carries less information, specifically in describing hand volume, hence leaving our method with a big disadvantage. However, despite this, the proposed model still outperforms the competitors on many metrics, and only marginally yields to the others. Additionally, we artificially introduced small Gaussian noise to the input 3D skeleton to simulate a network prediction; and we find that it has negligible impact on the model's accuracy.

### D. Qualitative evaluation

Fig. 5 illustrates the effect of the upscaling CNN on the NeRF output. Additionally, it demonstrates appearance transfer between different identities by providing corresponding CVAE embeddings to NeRF.

Qualitative comparisons of hand images rendered by the proposed approach to the *state-of-the-art* are demonstrated in Fig. 3 and Fig. 4. The figures show both the rendered hand and the ground truth image for visual comparison. Here we see excellent reproduction capability. There is some loss of detail, especially around the nails, and some smoothing. But on the whole, the results are visually very close to the natural images.

### E. Limitations

One of the limitations of our approach is a longer training time compared to hand parametric-based methods. The reason for this is that the parametric model reduces the problem's dimensionality, allowing the model to converge faster.

Additionally, extracting appearance and shape from a single image is a challenging task. We use CVAE latent embeddings that demonstrate good generalization, however, it may not always be accurate due to the single-view ambiguity.

## IV. CONCLUSIONS

We present a novel framework for 3D hand rendering that exploits a NeRF renderer that generalizes across multiple views and hand poses. The proposed method avoids the hard constraint of initialization and/or a parametric mesh model, widely adopted in the literature. Instead, we provide a step-by-step pipeline starting from point cloud extraction, and training of conditioned occupancy probabilities which are then combined into a NeRF as an implicit shape model to render 3D hands. The hand geometry is represented via occupancy probabilities and features, while appearance and shape are extracted and parametrized via a latent vector extracted from the image via a CVAE. The proposed NeRF conditioning combines these elements to efficiently render novel poses and views. On the benchmark publicly available INTERHAND2.6M dataset, our method achieves *state-of-the-art* accuracy.

## V. ACKNOWLEDGMENTS

This work was supported by the SNSF project 'SMILE II' (CRSII5 193686), the Innosuisse IICT Flagship (PFFS-21-47), EPSRC grant APP24554 (SignGPT-EP/Z535370/1) and and through funding from Google.org via the AI for Global Goals scheme. This work reflects only the author's views and the funders are not responsible for any use that may be made of the information it contains.

## ETHICAL IMPACT STATEMENT

In this paper, we used two publicly available benchmark hand datasets: InterHand2.6M [32] and HanCo [51, 52]. These datasets do not include any identifiable faces only images of hands, 3D keypoints, MANO mesh parameters, and segmentation masks. Additionally, the datasets are diverse in race and gender. We believe that using this data for quantitative and qualitative evaluation carries minimal risk of harm to participants originally captured in the datasets. We strictly follow the dataset protocols that use numerical identifiers to represent ground truth, maintaining privacy and respecting participant anonymity.

We do not anticipate any negative societal impacts from our research. The goal of this paper is solely to present a more efficient rendering method, which we believe will benefit research in computer vision. As with any artificial intelligence tool, there is a potential risk of the model being misused to blur the distinction between real and AI-generated hands. However, our intention is quite the opposite; by developing techniques for identity change using CVAE embeddings, our work aims to support anonymization. Moreover, in the provided visual figures, we demonstrate performance across different races to avoid potential bias or discrimination. This approach could reduce the need for real data collection, substituting it with rendering techniques that preserve privacy, which may help alleviate some of the ethical concerns associated with biometric data usage.

## REFERENCES


[1] T. Alldieck, H. Xu, and C. Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2021. 1

[2] M. Atzmon and Y. Lipman. Sal: Sign agnostic learning of shapes from raw data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[3] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5835–5844, 2021. 8

[4] Z. Cao, I. Radosavovic, A. Kanazawa, and J. Malik. Reconstructing hand-object interactions in the wild. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12397–12406, 2021. 2

[5] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017. 6

[6] P. Chen, Y. Chen, D. Yang, F. Wu, Q. Li, Q. Xia, and Y. B. Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12909–12918, 2021. 2

[7] X. Chen, B. Wang, and H.-Y. Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6, 7, 8

[8] E. Corona, T. Hodan, M. Vo, F. Moreno-Noguer, C. Sweeney, R. Newcombe, and L. Ma. Lisa: Learning implicit shape and appearance of hands. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6

[9] B. Deng, J. Lewis, T. Jeruzalski, G. Pons-Moll, G. Hinton, M. Norouzi, and A. Tagliasacchi. Neural articulated shape approximation. In *The European Conference on Computer Vision (ECCV)*. Springer, August 2020. 2

[10] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan. 3d hand shape and pose estimation from a single rgb image. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10825–10834, 2019. 1, 2

[11] Z. Guo, W. Zhou, M. Wang, L. Li, and H. Li. Handnerf: Neural radiance fields for animatable interacting hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21078–21087, June 2023. 2, 5, 6, 7, 8

[12] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 568–577, 2020. 1

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[14] Z. Huang, Y. Chen, D. Kang, J. Zhang, and Z. Tu. Phrit: Parametric hand representation with implicit template. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14974–14984, October 2023. 2

[15] B. Jiang, Y. Hong, H. Bao, and J. Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[16] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 7, 8

[17] J. M. Joyce. *Kullback-Leibler Divergence*, pages 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. 4

[18] K. Karunratanakul, S. Prokudin, O. Hilliges, and S. Tang. Harp: Personalized hand reconstruction from a monocular rgb video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[19] K. Karunratanakul, A. Spurr, Z. Fan, O. Hilliges, and S. Tang. A skeleton-driven neural occupancy representation for articulated hands. In *International Conference on 3D Vision (3DV)*, 2021. 2

[20] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, and S. Tang. Grasping field: Learning implicit representations for human grasps. *2020 International Conference on 3D Vision (3DV)*, pages 333–344, 2020. 1

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[22] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2022. 4

[23] D. Kulon, R. A. Guler, I. Kokkinos, M. M. Bronstein, and S. Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[24] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. 6

[25] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[26] M. Li, L. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu. Interacting attention graph for single image two-hand reconstruction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 4

[27] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2

[28] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2

[29] T. Luan, Y. Zhai, J. Meng, Z. Li, Z. Chen, Y. Xu, and J. Yuan. High fidelity 3d hand shape reconstruction via scalable graph frequency decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16795–16804, June 2023. 2

[30] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[31] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 5

[32] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 9

[33] A. Mundra, M. B R, J. Wang, M. Habermann, C. Theobalt, and M. Elgharib. Livehand: Real-time and photorealistic neural hand rendering, October 2023. 2, 5, 6, 7

[34] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[35] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Nerfies: Deformable neural radiance fields. *The IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[36] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10967–10977, 2019. 2

[37] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 7, 8

[38] R. A. Potamias, S. Ploumpis, S. Moschoglou, V. Triantafyllou, and S. Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4670–4680, June 2023. 2

[39] S. Prokudin, M. J. Black, and J. Romero. Smplpix: Neural avatars from 3d human models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1810–1819, 2021. 6

[40] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. 1

[41] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 1, 2

[42] A. Spurr, A. Dahiya, X. Wang, X. Zhang, and O. Hilliges. Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11230–11239, 2021. 1

[43] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[44] S.-Y. Su, F. Yu, M. Zollhöfer, and H. Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems*, 2021. 6

[45] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1

[46] S. Wang, A. Geiger, and S. Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[47] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022. 8

[48] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang. 3d hand pose tracking and estimation using stereo matching. *CoRR*, abs/1610.07214, 2016. 1

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[50] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng. End-to-end hand mesh recovery from a monocular rgb image. 2019. 1

[51] C. Zimmermann, M. Argus, and T. Brox. Contrastive representation learning for hand shape estimation. In *arxive*, 2021. 6, 8, 9

[52] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 6, 8, 9

[53] C. Zimmermann, D. Ceylan, J. Yang, B. C. Russell, M. Argus, and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 2, 4

[54] S. Zuffi, A. Kanazawa, D. W. Jacobs, and M. J. Black. 3d menagerie: Modeling the 3d shape and pose of animals. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5524–5532, 2016. 2