

OHTA: One-shot Hand Avatar via Data-driven Implicit Priors

Xiaozheng Zheng * Chao Wen * Zhuo Su Zeran Xu Zhaoju Li Yang Zhao Zhou Xue †
PICO, ByteDance

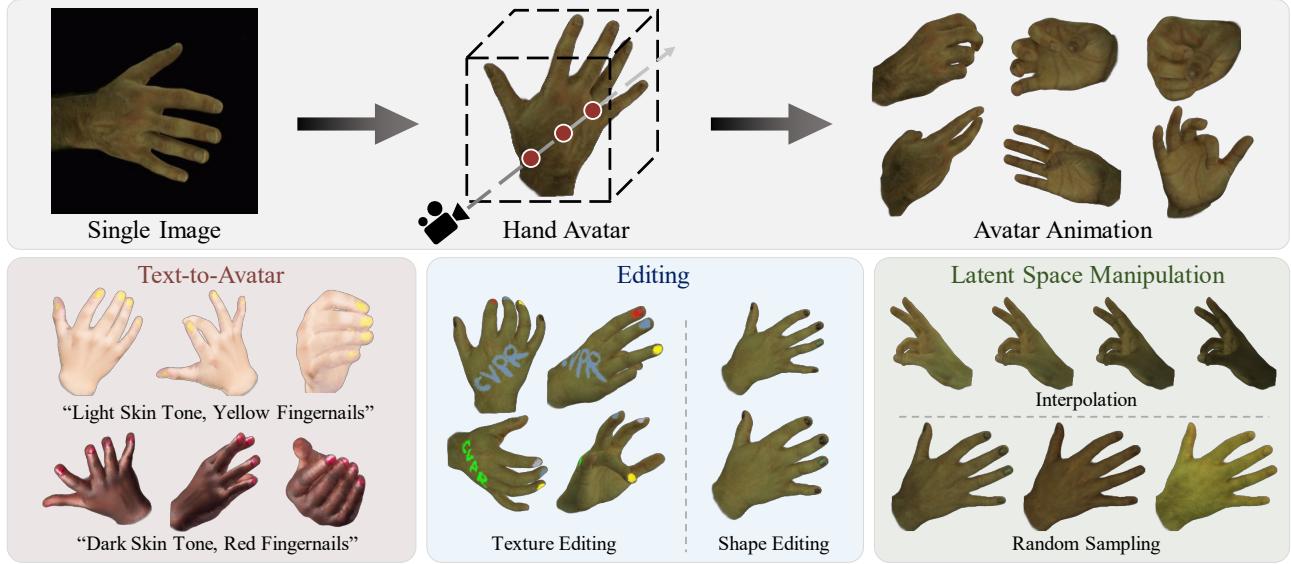


Figure 1. We introduce a novel approach capable of creating implicit animatable hand avatars using just a single image. Our framework facilitates 1) text-to-avatar conversion, 2) hand texture and geometry editing, and 3) interpolation and sampling within the latent space.

Abstract

In this paper, we delve into the creation of one-shot hand avatars, attaining high-fidelity and drivable hand representations swiftly from a single image. With the burgeoning domains of the digital human, the need for quick and personalized hand avatar creation has become increasingly critical. Existing techniques typically require extensive input data and may prove cumbersome or even impractical in certain scenarios. To enhance accessibility, we present a novel method OHTA (**One-shot Hand avaTAr**) that enables the creation of detailed hand avatars from merely one image. OHTA tackles the inherent difficulties of this data-limited problem by learning and utilizing data-driven hand priors. Specifically, we design a hand prior model initially employed for 1) learning various hand priors with available data and subsequently for 2) the inversion and fitting of the target identity with prior knowledge. OHTA demonstrates the capability to create high-fidelity hand avatars with consistent animatable quality, solely relying on a single image.

Furthermore, we illustrate the versatility of OHTA through diverse applications, encompassing text-to-avatar conversion, hand editing, and identity latent space manipulation.

1. Introduction

The progress in digital humans is reshaping our everyday lives by blending the physical and digital realms. In this emerging reality, human hands have a central role in creating a close and interactive experience, acting as the primary way people connect with the digital world. Consequently, it is vital to convert the hands of users into digital forms, allowing for the creation of personalized, controllable, and highly realistic representations in the virtual environment.

Conventional approaches of hand appearance modeling have typically relied on texture maps and colored meshes [10, 13, 14, 39, 60, 62]. However, the construction of personalized hand meshes and texture maps often demands the use of costly scanning data and artistic expertise. Recent efforts focus on creating hand avatars using data-driven approaches [9, 11, 16, 31, 55, 63]. Many of these studies have concentrated on creating neural im-

Project page: <https://zxx267.github.io/OHTA>

* Equal contribution, † Corresponding author

plicit hand avatars, as this form has demonstrated the ability to effectively model high-fidelity avatars across various human body parts [6, 9, 11, 23, 55, 75]. Despite the notable accomplishments in recent implicit hand representation, a noteworthy limitation persists, which entails the necessity of employing extensive sequential or multi-view images to procure a high-fidelity animatable hand avatar. This requirement poses a considerable challenge for general users, as the acquisition of such dense input data is often labor-intensive and, in certain cases, may not be practical. To broaden the scope of personalized hand avatar creation to more diverse situations, we aim to propose a novel approach capable of one-shot reconstruction of high-fidelity hand avatars. This endeavor presents a substantial challenge, given that it places personalized hand avatar creation at a crossroads, necessitating a data-driven method despite the scarcity of available data.

In response to this challenge, we introduce OHTA (**O**ne-shot **H**and **a****T****A**r) which captures data-driven priors to facilitate one-shot high-fidelity hand avatars creation. Our motivation stems from the fact that leveraging priors can complement unobserved information due to the inherent similarities shared by diverse human hands. To this end, we decouple our framework into two stages: 1) hand prior knowledge learning, and 2) one-shot reconstruction with the aid of prior knowledge. In the first stage, we employ the Hand Prior Network (HPNet) to capture prior knowledge from the accessible data with multiple identities. In the second stage, we utilize HPNet to inverse a similar hand and subsequently optimize it to the target identity with prior knowledge for regularization, based on the given image.

HPNet plays a pivotal role since its capacity for both 1) *prior transferring* and 2) *prior learning* significantly impacts the ultimate one-shot performance. Therefore, designing such a network is non-trivial and presents a dual challenge: the architecture must be well-suited for two purposes simultaneously. To address the challenges, we adopt corresponding designs in HPNet. For *prior transferring*, we design HPNet in a mesh-guided manner since the aid of the mesh enables the learned prior knowledge to transfer to novel identities. Both geometry and texture priors learning of HPNet are guided by the mesh information, achieving robustness towards novel poses and identities. For *prior learning*, we equip HPNet with 1) the multi-resolution field based on the mesh scaffold and 2) identity-specific albedo and identity-shared shadow disentanglement. The multi-resolution field facilitates learning detailed texture prior, and the disentanglement models transferable self-occlusion effect, both of which contribute to high-fidelity one-shot hand avatars. With all these designs, OHTA is capable of addressing many downstream tasks (as shown in Fig. 1), including 1) hand avatar creation using text prompts, 2) editing of hand geometry and appearance, and 3) appearance

interpolation and sampling in identity latent space.

In summary, our contributions can be listed as follows:

- We introduce the first framework for one-shot implicit hand avatar creation. These avatars exhibit high-fidelity and consistent animatable quality.
- We present the Hand Prior Network exploiting transferable geometry, albedo, and shadow priors, contributing to the high fidelity of one-shot hand avatars.
- We substantiate the efficacy and robustness of our framework through a comprehensive series of experiments and showcase its utility in diverse applications, including text-to-avatar, hand editing, and latent space manipulation.

2. Related Work

Animatable Hand Avatar. Research in creating personalized, animatable hand avatars has advanced through two main approaches. Some works utilize explicit geometry and texture. HTML [62] is prominent for its texture inference on the MANO model [65]. NIMBLE [39] extends the hand skeleton model [38], capable of inferring surface with texture maps. Handy [60] improves realism using a GAN-based texture model [29], but these methods often suffer from poor adaptability due to limited training data. HARP [31] addresses this by adjusting geometry and texture for individual identities. Another line of research explores neural implicit representations for more realistic texture modeling. LISA [11] learns an implicit color and surface representation. HandAvatar [9] employs disentangled implicit representations for hand geometry, albedo, and illumination. HandNeRF [16] and LiveHand [55] extend NeRF [51] techniques with deformation fields for hands, even achieving real-time performance. HO-NeRF [63] further explores hand-object reconstruction using implicit avatars. Notwithstanding their accomplishments, it is worth noting that these models are unable to create high-fidelity hand avatars from a single image, as is achievable with our proposed approach.

One-shot Human Avatar. A similar task to one-shot hand avatar modeling is reconstructing a human body avatar from a single image, with methods divided into “explicit” and “implicit” categories. Explicit methods [1, 2, 5, 27, 28, 34–36, 47, 57, 85] rely on parametric models [46, 56, 57] to estimate a minimally-clothed body or to regress body shape offsets, often struggling with loose clothing representation. DINAR [70] integrates neural textures with the explicit model for enhanced photo-realism. Implicit methods, using topology-agnostic fields, aim for higher realism. Many works [17, 66, 67, 83] utilize large datasets to estimate body avatars with pixel-aligned features, while others [18, 21] further propose avatar animation. However, these approaches are often constrained by training data, limiting their generalizability. NeRF-based methods [24, 58, 69, 75, 77, 82] typically require dense inputs but have evolved to accommodate few-shot or one-shot sce-

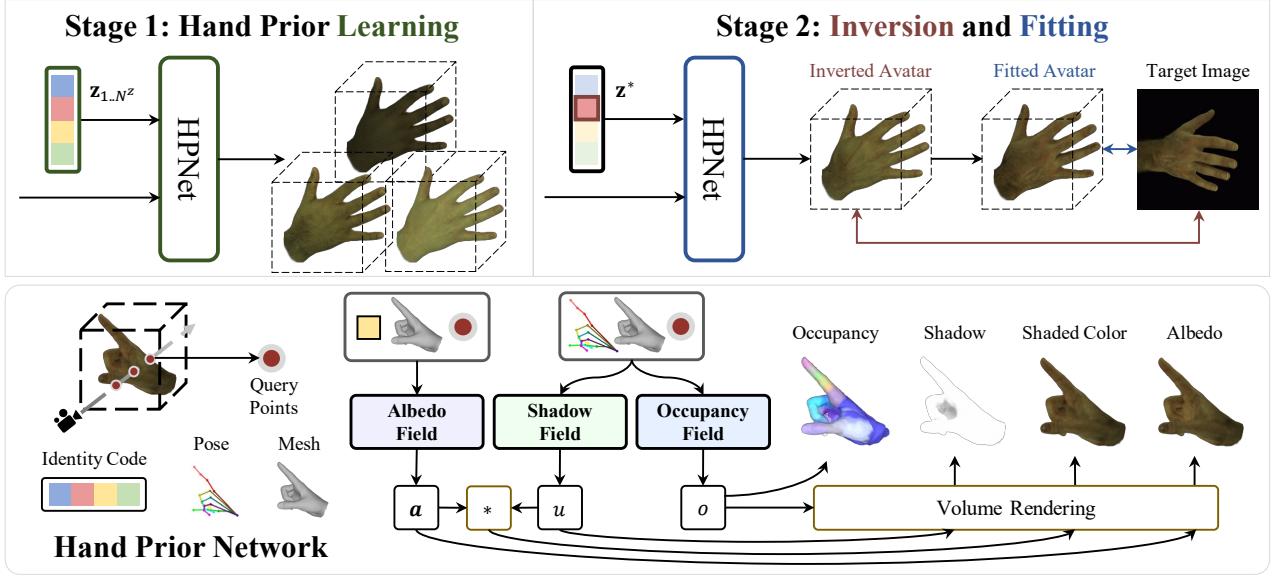


Figure 2. The two-stage framework of OHTA (above) and the Hand Prior Network (below). For stage-1, OHTA optimizes identity code and HPNet to capture various hand priors. For stage-2, OHTA first optimizes identity code for texture inversion, then optimizes HPNet for texture fitting to capture the details. HPNet consists of three fields (*i.e.* albedo, shadow, and occupancy) for capturing transferable hand prior knowledge. Combined with the albedo, shadow, and occupancy values, we use volume rendering to obtain final shaded color images.

narios by incorporating various priors. ActorsNeRF [54] adapts to new subjects by pre-training category-level human NeRFs, while ELICIT [20] and SHERF [19] employ semantic priors or 3D-aware features for one-shot avatar creation. For faces, CG-NeRF [26] supports one-shot 3D-aware face synthesis with a conditional NeRF. PhoneScan [7] learns identity and expression prior with CNNs from large-scale data, and finetunes the prior model for few-show head avatar creation. Preface [6] also trains an identity-conditioned prior model and achieves few-shot high-fidelity 3D head creations with its inversion and fitting. Compared with PhoneScan and Preface, our method further exploits geometry, albedo, and shadow priors to create one-shot animatable hand avatars with consistent driving performance.

Mesh-guided Representation. Parametric models generate meshes from pose and shape parameters [37, 46, 57, 65], but these meshes have unalterable topology structure. Implicit representations are increasingly researched to address these limitations, with studies focusing on human [3, 8, 12, 49, 50, 68, 71, 72] and hand geometry [9, 22, 30], often using rigid transformations or skeletons for prediction. COAP [50] and PairOF [9] enhance robustness by incorporating mesh information and localized encoders or part-pair-wise decoders, respectively. Mesh information also aids implicit texture modeling [4, 9, 45, 59, 76, 78, 84]. NeuMesh [78] and DE-NeRF [76] place features on vertices for editable radiance fields. Recent implicit human body [59, 84] and face [4, 15] modeling also rely on mesh information for guidance. HandAvatar [9] uses barycentric sampled anchor upon meshes for hand modeling. Our

method differs by employing multi-resolution fields on the mesh scaffold for enhanced robustness and fidelity.

Single Image to 3D. Recently, there has been a surge in research focused on learning how to create 3D models from a single image [48, 79]. Zero-1-to-3 [43] pioneers the zero-shot single-image to 3D conversion. Recent works like One-2-3-45 [42], Magic123 [61], and Consistent123 [41] build upon Zero-1-to-3 to obtain more 3D-consistent results by introducing more priors. Although such methods have pre-trained networks on large datasets, there is an underutilization of specific domain knowledge. In contrast to these methods, we are dedicated to considering the distinctive properties of the hand to create animatable avatars.

3. Method

In this section, we provide a detailed description of the proposed hand representation and complete avatar reconstruction pipeline (Fig. 2). We first introduce a novel hand representation that encodes multi-resolution neural implicit fields at a mesh scaffold (Sec. 3.1). As for the one-shot hand avatar creation pipeline, we decouple it into two stages. In the first stage (Sec. 3.2), we devise an effective training scheme that captures hand prior knowledge through the HPNet. In the second stage (Sec. 3.3), we invert and optimize the HPNet to create high-quality hand avatars according to a single input image.

3.1. OHTA Hand Representation

Our representation of the hand comprises implicit geometry and neural texture fields based on a mesh scaffold.

Mesh Scaffold. We adopt MANO-HD [9], a super-resolution variant of MANO [65], as the mesh scaffold. (1) For training HPNet, we use personalized MANO-HD by leveraging a multi-layer perceptron (MLP) \mathcal{M}_{shape} to refine the template mesh. The refined template mesh is defined as $\bar{\mathbf{M}}^+ = \bar{\mathbf{M}} + \mathcal{M}_{shape}([\mathcal{P}(\bar{\mathbf{M}}), \theta, \mathbf{z}])$, where $\bar{\mathbf{M}}$, \mathcal{P} , θ , \mathbf{z} and $[\cdot]$ denote the template mesh, positional encoding, pose parameters, identity code and concatenation respectively. The MLP can be optimized via IoU loss \mathcal{L}_{shape} , detailed in [44]. The template mesh can be deformed to posed mesh with a Linear Blend Skinning transformation according to pose parameters. (2) For one-shot hand avatar reconstruction, where acquiring a precisely refined mesh from a single view is unfeasible, we use a mesh \mathbf{M} represented by shape β and pose θ parameters as the input. For simplicity, the input mesh is denoted as \mathbf{M} in the following text.

Geometry Network. To achieve robust geometry modeling towards novel identities and poses, we utilize PairOF [9] to predict occupancy values. Given N^q query points $\mathbf{q} \in \mathbb{R}^{N^q \times 3}$, hand mesh and hand pose parameters, PairOF $f^\mathcal{O}$ predicts the occupancy value $o_{\mathbf{q}} = f^\mathcal{O}(\mathbf{q}, \mathbf{M}, \theta)$ to indicate whether the point locates inside or out of the surface. Hence, the hand surface can be formulated as $\{\mathbf{q} | o_{\mathbf{q}} = 0.5\}$. For more details, please refer to [9, 50].

Multi-resolution Field. To achieve high-fidelity texture modeling, we designed a Multi-resolution Field (Fig. 3) with the mesh guidance. More specifically, we uniformly sample N_k^p points \mathbf{P}_k on the input mesh and represent them with barycentric coordinates for each resolution, where $k \in \{1, \dots, K\}$ denotes the resolution level. For each point in \mathbf{P}_k , we attach them with a learnable point encoding of N^c dimension that is randomly initialized. In this way, we can obtain point encodings \mathbf{E}_k of a resolution. For query points \mathbf{q} , we conduct spatial interpolation to acquire the queried encoding of this resolution $\mathbf{Q}_k = \text{interp}(\mathbf{q}, \mathbf{E}_k)$. Specifically, we first extract the encodings of N^n neighbor points $\mathcal{K}(\mathbf{E}_k) \in \mathbb{R}^{N^n \times N^q \times N^c}$ of \mathbf{P}_k , where \mathcal{K} denotes k -nearest neighbors. Then, we perform a weighted average with the inverse Euclidean distances of those points to the query point as the weights to acquire the queried encoding of this resolution $\mathbf{Q}_k \in \mathbb{R}^{N^q \times N^c}$. After that, we feed queried encodings with the field-specific information ψ to resolution-specific MLP \mathcal{M}_k to obtain resolution-specific features: $\mathbf{D}_k = \mathcal{M}_k(\mathbf{Q}_k, \psi)$, where $\mathbf{D}_k \in \mathbb{R}^{N^q \times N^d}$, and ψ is designed to distinguish the albedo-specific field $f^\mathcal{A}$ and shadow-specific field $f^\mathcal{U}$, which are combined with the multi-resolution mechanism to better represent the texture, as described later. Finally, we make use of another MLP \mathcal{M}_{fuse} to fuse resolution-specific features of multi-resolutions to predict the target field value $\mathbf{x}_q = \mathcal{M}_{fuse}([\mathbf{D}_1 \dots, \mathbf{D}_K])$. The overall hand representation comprises the above components, as defined formally:

$$\mathcal{H} : (\mathbf{q}, \mathbf{M}, \theta, \mathbf{z}; f^\mathcal{A}, f^\mathcal{U}, f^\mathcal{O}) \mapsto (o, c), \quad (1)$$

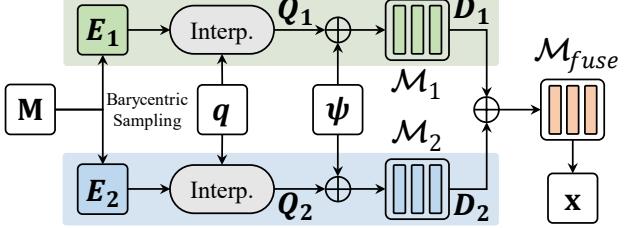


Figure 3. Multi-resolution Field. For simplicity, we take two resolutions for example. \oplus denotes feature concatenation.

where o and c denote the occupancy and shaded color.

3.2. Hand Prior Learning

To exploit the potential of OHTA representation for one-shot reconstruction, learning transferrable and abundant hand prior knowledge is of priority. Inspired by [7], we propose a novel paradigm for encoding those prior to the HPNet through a set of accessible data from multiple identities. On the dataset including N^z identities, our goal is to learn multiple hand avatars within a network, with the respective identity codes $\mathbf{z}_{1\dots N^z}$ and network parameters optimized.

Geometry Prior. For the hand geometry prior learning, we first pre-train the PairOF module (*i.e.* occupancy field) $f^\mathcal{O}$. We utilize data from different hand shapes to make HPNet learn the geometric priors of the hand which can transfer to new identities. With hand meshes derived from input \mathbf{M} , we sample point clouds with N^o points as training data [50].

Texture prior. Subsequently, we focus on learning the texture prior knowledge of the hand. We observe that the albedo features are specific to each identity, while the shadows, which arise due to the hand’s self-occlusion, are common across different identities. Hence, we disentangle the texture prior learning into albedo and shadow prior learning. Our methodology leverages a multi-resolution field representation for simultaneous learning of both albedo and shadow fields by incorporating field-specific information ψ .

Identity-specific Albedo Field. Since the albedo of the hands is dependent on the identity other than pose etc., we set $\psi = \mathbf{z}$ for identity-specific albedo field modeling. For query points \mathbf{q} , we obtain the albedo $\mathbf{a}_q = f^\mathcal{A}(\mathbf{q}, \mathbf{M}, \mathbf{z})$.

Identity-shared Shadow Field. Modeling identity-specific environmental lighting conditions of hands like HandAvatar [9] cannot capture shadow prior knowledge that can generalize to one-shot reconstruction. To overcome this problem, our approach aims to model a more universal shadow prior, grounded in the observation that shadows from self-occlusion are largely consistent across different hands when performing identical poses. Hence, we build the shadow field with $\psi = \theta$, where θ is pose parameters without global rotations. For query points \mathbf{q} , we obtain the shadow $u_q = f^\mathcal{U}(\mathbf{q}, \mathbf{M}, \theta)$.

Combining the albedo field and shadow field, we ob-

tain a complete Multi-resolution Texture Field (MTField). For query points \mathbf{q} , we derive the shaded color value $\mathbf{c}_\mathbf{q} = u_\mathbf{q} \mathbf{a}_\mathbf{q} \in \mathbb{R}^{N^q \times 3}$. For a casted camera ray \mathbf{r} , we uniformly sample N^q queries $\{\mathbf{q}_i\}_{i=1}^{N^q}$, each of which has occupancy o and texture \mathbf{c} . We render each pixel $\mathbf{I}_\mathbf{r}$ of image via volume rendering [51]:

$$\mathbf{I}_\mathbf{r} = \sum_{i=1}^{N^q} \left(\prod_{j=1}^{i-1} (1 - o_{\mathbf{q}_j}) \right) o_{\mathbf{q}_i} \mathbf{c}_{\mathbf{q}_i}. \quad (2)$$

Optimization. The training of HPNet consists of two steps. In the first step, we pre-train $f^\mathcal{O}$ with the MANO pose parameters annotations and shape parameters randomly sampled from a given range. The objective for optimizing PairOF is to minimize the error between ground truth o^* and predicted occupancy values, *i.e.*, $\mathcal{L}_{occ} = \frac{1}{N^o} \sum_{\mathbf{q}} (o_{\mathbf{q}} - o_{\mathbf{q}}^*)^2$. In the second step, we conduct end-to-end training for the mesh scaffold, occupancy field, and MTField. Specifically, MTField is optimized with shadow regularization $\mathcal{L}_{shadow} = \|\mathbf{U} - \mathbf{1}\|_1$ and the reconstruction loss function (consisting of LPIPS loss [81] and l_1 loss):

$$\mathcal{L}_{rec} = \mathcal{L}_{LPIPS}(\mathbf{I}, \mathbf{I}^*) + \|\mathbf{I} - \mathbf{I}^*\|_1, \quad (3)$$

where \mathbf{U} , \mathbf{I} and \mathbf{I}^* denote the shadow values, the rendered image, and the ground truth image. The shadow field is regularized to be close to 1 to ensure that it focuses more on the shadow caused by self-occlusion. The full loss is a combination of the previous terms: $\mathcal{L} = \mathcal{L}_{shape} + \mathcal{L}_{occ} + \mathcal{L}_{rec} + \lambda_{shadow} \mathcal{L}_{shadow}$, where λ_{shadow} is used to balance different loss terms.

3.3. One-shot Hand Avatar Reconstruction Pipeline

After hand prior learning, HPNet encodes various prior knowledge useful for one-shot reconstruction. For each input image, we only optimize part of HPNet to reconstruct the target hand avatar. The complete pipeline is as follows.

Preprocessing. We utilize the off-the-shelf hand pose estimator [64] to predict the shape parameter β , pose parameter θ , and camera pose of the input images.

Geometry. Since it is impractical to learn a refined mesh based on an image, we adopt meshes derived from hand shape parameters β as the input for occupancy prediction.

Texture Inversion. To provide good prior knowledge for the one-shot reconstruction, we first inverse a similar appearance from the MTField. More specifically, we solve an optimization problem to find identity code \mathbf{z}^* and per-channel color calibration coefficients $\{\mathbf{w}, \mathbf{b}\}$ that can produce a similar appearance to the target identity of the input image. The color calibration alleviates the hand skin bias caused by the data used for prior learning, assisting better texture inversion for the target identity. It works as follows: $\tilde{\mathbf{a}}_\mathbf{q} = \mathbf{w} * \mathbf{a}_\mathbf{q} + \mathbf{b}$. During the inversion optimization, we keep the network weights frozen. Given an input image \mathbf{I}^* of the target identity, we optimize to render an image \mathbf{I} that is similar to the target identity using our MTField with \mathbf{z}^*

and $\{\mathbf{w}, \mathbf{b}\}$. This procedure is optimized with the reconstruction loss function in Eq. (3): $\arg \min_{\mathbf{z}^*, \mathbf{w}, \mathbf{b}} \mathcal{L}_{rec}$.

Texture Fitting. The goal of the texture fitting is to adapt the weights of the MTField to capture the details of the target identity from the input image. We utilize prior knowledge in this procedure by two means. First, we only fine-tune the resolution-specific MLPs $\{\mathcal{M}_k\}_{k=1}^K$ and texture feature fusing MLP \mathcal{M}_{fuse} of albedo field f^A , while keeping other components (*e.g.*, the point encodings and shadow field) frozen. Hence, we can transfer the prior knowledge (*e.g.*, shadow prior) well to the target avatar. In this stage, all learnable parameters are denoted as ξ . Second, we apply view regularization to avoid overfitting the target view. Specifically, we constraint the texture-fitting results of some reference views with different poses $\{\mathbf{R}_i\}_{i=1}^{N^r}$ to be close to the rendering results before texture-fitting $\{\tilde{\mathbf{R}}_i\}_{i=1}^{N^r}$, where N^r is the number of the generated reference view. With the prior knowledge, our reconstructed hand avatar can achieve stable animation while preserving the details of the target identity. This fitting is conducted by minimizing a series of the reconstruction loss function in Eq. (3):

$$\arg \min_{\xi} \mathcal{L}_{fit} = \mathcal{L}_{rec}(\mathbf{I}, \mathbf{I}^*) + \lambda_{ref} \sum_{i=1}^{N^r} (\mathcal{L}_{rec}(\mathbf{R}_i, \tilde{\mathbf{R}}_i)), \quad (4)$$

where λ_{ref} is used to balance different loss terms.

4. Experiment

4.1. Implementation Details and Metrics

Pre-training of PairOF. We adopt all right-hand annotations in InterHand2.6M [53] for pre-training like HandAvatar [9]. To make the occupancy field more robust to new identities, we randomly sample the hand shape parameters β from a range $\pm 3\sigma$ ensuring that the hand shapes are physically plausible. We set $N^o = 256$ for training with \mathcal{L}_{occ} .

End-to-end Prior Learning of HPNet. We utilized 21 subjects from the InterHand2.6M [53] training set as the training data for HPNet. Specifically, each subject used one capture and left out 7 unseen poses for evaluation. For the albedo field, we set $K = 4$ and $N_k^p = \{512 \times 2^{k-1}\}_{k=1}^4$. For the shadow field, we set $K = 1$ and $N_1^p = 256$. In addition, we set $N^n = 4$, $N^q = 64$, $N^c = 128$, and $N^d = 16$. The identity code \mathbf{z} is learnable parameters initialized from a truncated normal distribution with standard deviation $\sigma = 0.02$. The rendering resolution is 256×256 .

One-shot Reconstruction. We use an off-the-shelf estimator [64] to predict the MANO parameters of the in-the-wild images. We set $N^r = 7$ and $\lambda_{ref} = 0.2$. The reconstruction first takes 50 steps for texture inversion and then 100 steps for texture fitting (56 minutes with an A100 GPU). For more details, please refer to our supplementary materials.

Metrics. Consistent with previous works [9, 11, 55], we

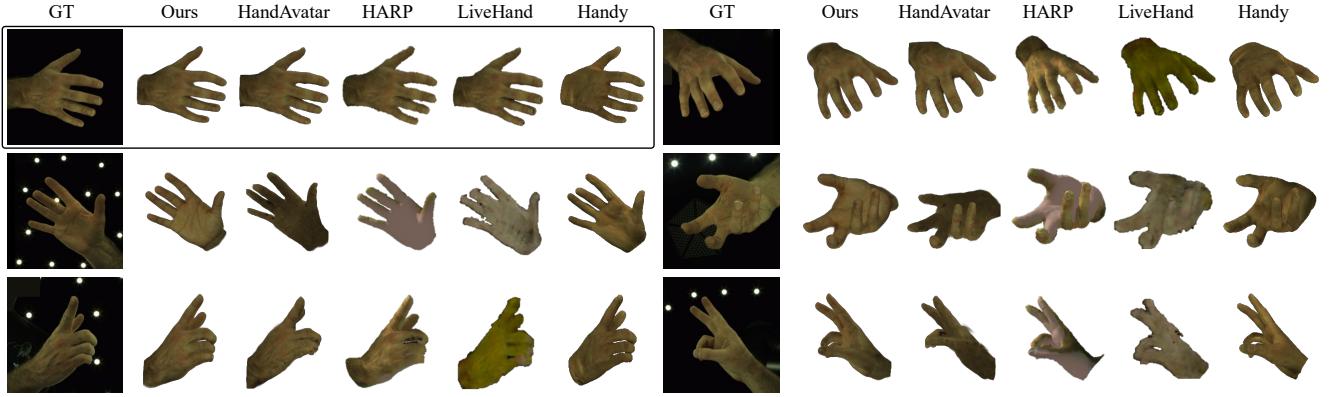


Figure 4. Qualitative comparison with state-of-the-art methods on InterHand2.6M [53]. The black box indicates the input image.

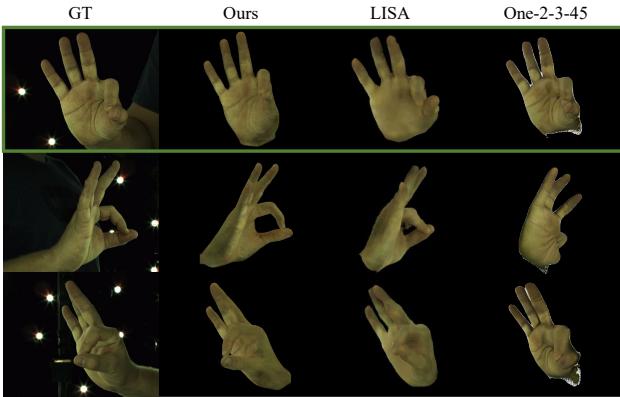


Figure 5. Comparison of novel view synthesis on InterHand2.6M [53]. The green box indicates the input view.

report LPIPS [81], PSNR, and SSIM [74] to reflect image similarity as the metrics of rendering quality.

4.2. Evaluation of One-shot Avatar Reconstruction

We comprehensively compare OHTA with previous methods [9, 31, 55, 60] under the one-shot reconstruction scenario with different data sources. Additionally, we benchmark against methods [9, 23, 75] that employ a large number of images, aiming to gauge the potential performance ceiling in one-shot scenarios. Moreover, we conduct novel-view comparisons with LISA [11] and One-2-3-45 [42].

InterHand2.6M. We conduct evaluations on the testing set of InterHand2.6M [53], following HandAvatar [9]. The results of training on multi-view sequences are reported in [9]. The results of training with a single image (*i.e.* one-shot) are based on their open-source codes [9, 31, 55, 60]. We do not update the hand pose parameters for HARP since the official implementation for a single image may collapse when updating them. As for Handy [60], we begin by converting the MANO labels into Handy’s format. We then optimize the latent vectors of the texture model in Handy and fit the single image using differentiable rendering. As shown in

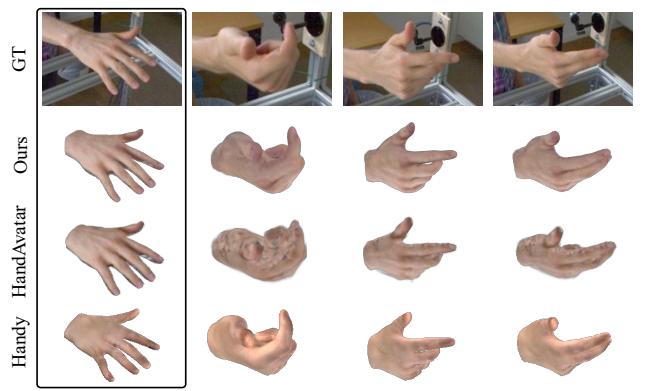


Figure 6. Qualitative comparison on HanCo [86].

Tab. 1, our one-shot performance is significantly better than the existing methods [9, 31, 55, 60]. Note that # denotes the number of images. The qualitative comparisons in Fig. 4 demonstrate our robust performance for novel views and poses. All the results demonstrate that data-driven priors can benefit the one-shot performance remarkably. Meanwhile, our performance is comparable to that of methods using all the images for training, which further validates the quality of the avatars.

Fig. 5 presents the novel view synthesis comparisons with LISA [11] and One-2-3-45 [42]. Due to the unavailability of models, we use results reported in the original LISA paper for comparison. Our one-shot performance can capture more details than LISA, owing to our adequate prior knowledge from the prior model. Due to the underutilization of specific domain knowledge, the general 3D generation approach (One-2-3-45) cannot generate plausible novel views for animatable hand avatar creation.

HanCo. To show OHTA is robust for the appearance different from the training data of the HPNet, we also compare with existing methods on the HanCo [86] dataset. We take HandAvatar [9] and Handy [60] for comparisons since they show robust performance for one-shot reconstruction on In-

Method	#Train	PSNR↑	LPIPS↓	SSIM↑
SelfRecon [23]	11,757	26.38	14.21	0.879
HumanNeRF [75]	11,757	27.64	11.45	0.884
HandAvatar [9]	11,757	28.23	10.35	0.894
HandAvatar [9]	1	23.79	17.78	0.820
HARP [31]	1	19.82	22.49	0.761
LiveHand [55]	1	23.01	18.64	0.763
Handy [60]	1	25.56	14.98	0.794
Ours	1	26.11	12.93	0.864

Table 1. Evaluation results on InterHand2.6M.

Method	PSNR↑	LPIPS↓	SSIM↑
HandAvatar [9]	19.76	14.41	0.846
Handy [60]	21.11	11.71	0.768
Ours	22.15	11.55	0.886

Table 2. Evaluation results on HanCo.

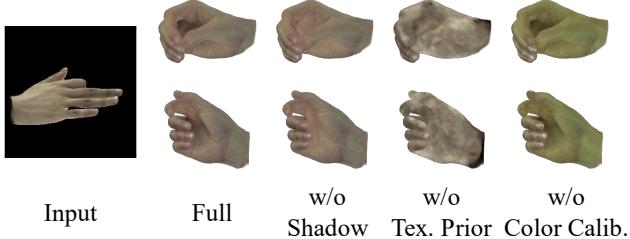


Figure 7. Visual results for ablation study with in-the-wild input.

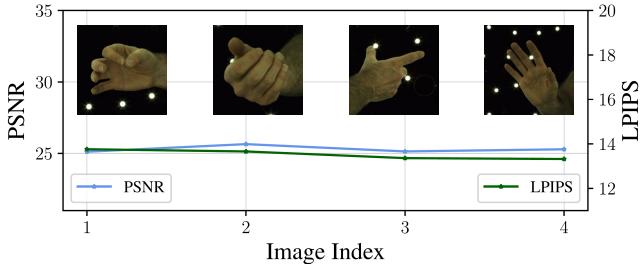


Figure 8. Robustness towards different input images. The images above the figure are the corresponding input images.

terHand2.6M (Tab. 1). We take 1 sequence of 4 cameras for the experiment, with 1 frame for training and others for testing. For more details, please refer to our supplementary materials. As shown in Tab. 2, our method outperforms other methods consistently in all metrics. We also show the qualitative comparisons in Fig. 6, demonstrating our method can obtain consistent performance for free-pose animation.

4.3. Ablation Study

Our ablation studies contain two parts. One part is for justifying our components of HPNet. The other part is to show our designs are effective for one-shot reconstruction.

Method	PSNR↑	LPIPS↓	SSIM↑
Full Model	27.64	12.23	0.896
w/o Multi-resolution	27.11	12.96	0.890
w/o Shadow Field	27.09	12.62	0.889
w/o Shape Fitting	27.27	12.63	0.892

Table 3. Ablation of HPNet for unseen poses.

#	Method	PSNR↑	LPIPS↓	SSIM↑
1	Full Model	26.11	12.93	0.864
2	w/o Tex. Prior	26.01	14.35	0.863
3	w/o Tex. Fitting	25.87	13.11	0.863
4	w/o Inversion	25.47	13.26	0.860
5	w/o Color Calib.	26.05	13.09	0.862
6	w/o Regularization	25.76	13.56	0.857

Table 4. Ablation of one-shot reconstruction on InterHand2.6M.

Hand Prior Network. The evaluation is conducted in 7 unseen poses as described in Sec. 4.1. As shown in Tab. 3, employing multi-resolution fields contributes greatly to all the metrics. Meanwhile, modeling pose-aware, identity-shared shadow is beneficial for overall performance, validating the existence of a shared self-occlusion effect for hands. At last, we demonstrate learning identity-specific hand shape offset is significantly important when the data is abundant to remove the shape inaccuracy of MANO fitting.

One-shot Reconstruction. We show the effectiveness of our strategies for one-shot reconstruction in Tab. 4 (on InterHand2.6M as Sec. 4.2) and Fig. 7 (in-the-wild). When there is no texture prior (#2), a significant drop in LPIPS is observed, indicating that details of the hands are missing (Fig. 7). When omitting texture fitting (#3), the performance slightly drops, indicating that our inversion strategy can optimize similar hands from the latent space when the training data is adequate. When not performing inversion (setting the identity features are zero vectors, #4), the results become worse. This further shows that inversion can provide better prior knowledge for the one-shot reconstruction. When ignoring color calibration (#5), the performance will degenerate even with the data captured under the same camera setups. When using in-the-wild data for reconstruction (Fig. 7), adopting color calibration is of more significance. We also demonstrate that overfitting the input image without view regularization leads to poorer quality (#6). Finally, we qualitatively show that our self-occlusion modeling with a shadow field can contribute to more realistic hand avatar animation (Fig. 7).

4.4. Robustness towards Diverse Inputs

To show the robustness of OHTA, we present our quantitative and qualitative performance across diverse inputs.

Quantitative. As depicted in Fig. 8, OHTA gets similar

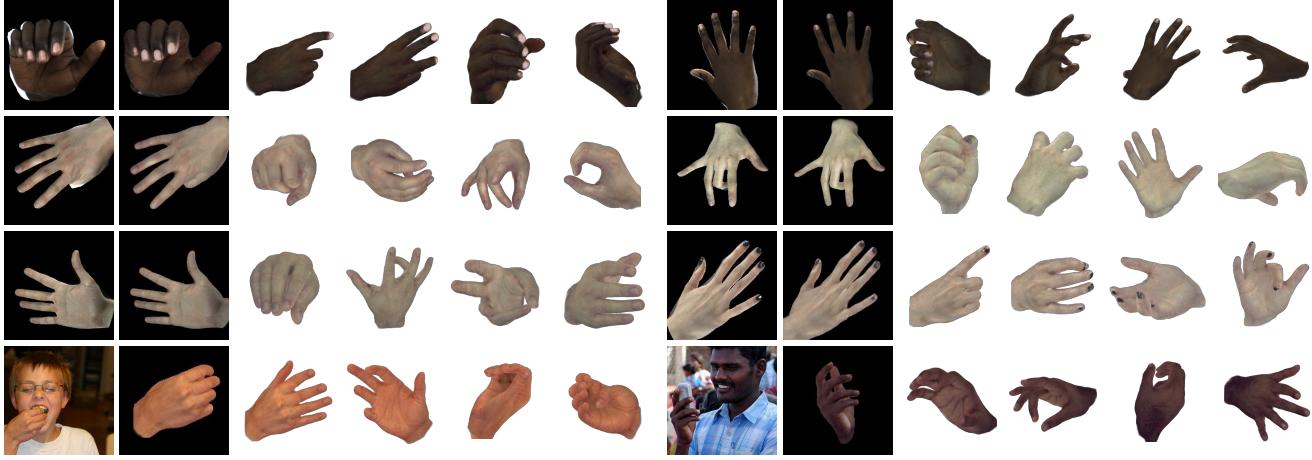


Figure 9. In-the-wild results from real-captured images and MSCOCO [40] dataset. The last line shows results on MSCOCO.

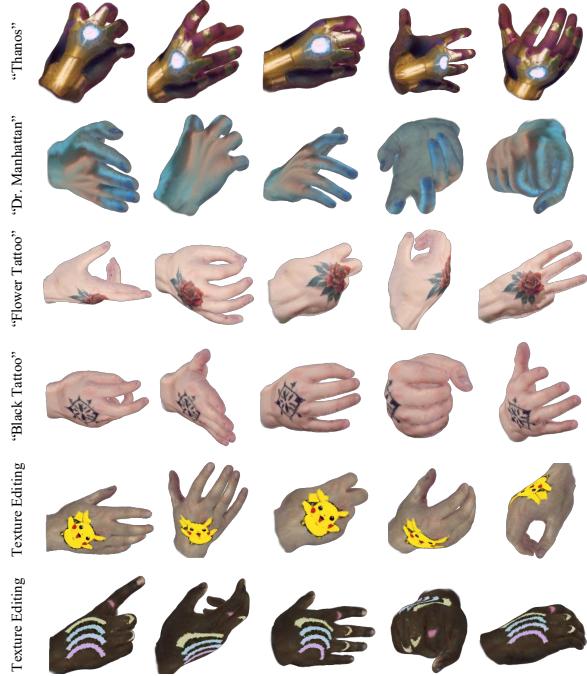


Figure 10. Text-to-avatar and texture editing.

PSNR and LPIPS for different inputs on InterHand2.6M.

Qualitative. In Fig. 9, we illustrate our results on various real-captured images and whole-body images of the MSCOCO dataset [25, 40]. OHTA performs consistently across different skin tones, hand poses, and viewpoints for real-captured data. Also, OHTA can reconstruct hand avatars from human images on MSCOCO, which further justifies the robustness of OHTA. We refer the reader to the supplementary material for additional qualitative results.

4.5. Application

Based on our one-shot framework, we can achieve 1) hand avatar creation with text prompts [80], 2) hand avatar edit-

ing of the geometry and appearance, and 3) identity-space manipulation including appearance sampling and interpolation. The results are shown in Fig. 1 and Fig. 10. For more details, please kindly refer to our supplementary materials.

Text-to-avatar. We use ControlNet [80] to generate a hand image with the hand mask and text prompts, then utilize OHTA to reconstruct hand avatars from the hand images.

Editing. The geometry is based on the mesh scaffold. Thus, we can edit the hand shape parameters β to edit the geometry. For appearance editing, we can draw arbitrary content on a target view of the hand avatar and then use OHTA to update the edited content part with its mask.

Latent space manipulation. Our identity latent space is continuous. Hence, we can sample the identity code from a normal distribution to sample hand avatars. Furthermore, we can interpolate two target identity codes to obtain smooth appearance interpolations between two identities.

5. Conclusion

In this paper, we present the first one-shot implicit hand avatar creation approach, OHTA. OHTA consists of two stages. The first stage focuses on learning the hand prior on datasets with multiple identities. For the creation of one-shot hand avatars, it is only necessary to optimize the second stage with the learned prior, which includes texture inversion and texture fitting. OHTA is robust for diverse input images and is capable of solving various downstream tasks with consistent animations. These tasks include hand avatar creation with text prompts, hand geometry and appearance editing, and manipulation of identity latent space.

Limitations and Future Works. OHTA exhibits suboptimal performance for input images that feature 1) notably uneven lighting, and 2) highly inaccurate pose estimation. Thus, further designing the approach to be more robust for those situations is worthy of ongoing exploration.

OHTA: One-shot Hand Avatar via Data-driven Implicit Priors

Supplementary Material

In the supplemental material, we provide:

- § A: Implementation Details.
- § B: More Experiments and Results.
- § C: Discussions.
- § D: More Qualitative Results.

A. Implementation Details

A.1. Spatial Interpolation

We uniformly sample N_k^p anchor points \mathbf{P}_k on the input mesh and represent them with barycentric coordinates for each resolution, where $k \in \{1, \dots, K\}$ denotes the resolution level. The sampled barycentric coordinates are fixed after the sampling, which means the sampling for each resolution is only conducted once for obtaining the fixed point encodings \mathbf{E}_k . That means that \mathbf{P}_k are 3D points representing the location of \mathbf{E}_k .

For query points \mathbf{q} , we conduct spatial interpolation to acquire the queried encoding of this resolution:

$$\mathbf{Q}_k = \text{interp}(\mathbf{q}, \mathbf{E}_k). \quad (5)$$

Fig. A depicts the spatial interpolation process. Specifically, we first extract the encodings of N^n neighbor points $\mathcal{K}(\mathbf{E}_k) \in \mathbb{R}^{N^n \times N^q \times N^c}$ of \mathbf{P}_k , where \mathcal{K} denotes k -nearest neighbors. Then, we perform a weighted average with the inverse Euclidean distances of those points to the query point as the weights (values indicated by the **line colors** in Fig. A) to acquire the queried encoding of this resolution $\mathbf{Q}_k \in \mathbb{R}^{N^q \times N^c}$ (*i.e.* the features of the **red point** in Fig. A).

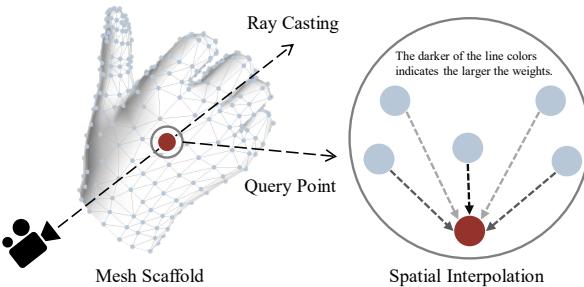


Figure A. Illustration of spatial interpolation for a level of resolution.

Furthermore, to illustrate the impact of different resolutions on spatial interpolation, we show the anchor points \mathbf{P}_k used in multi-resolution fields at different resolutions in Fig. B. The light red area in the figure indicates the region used for interpolation to obtain \mathbf{Q}_k . For more discussion on multi-resolution, please refer to Sec. B.

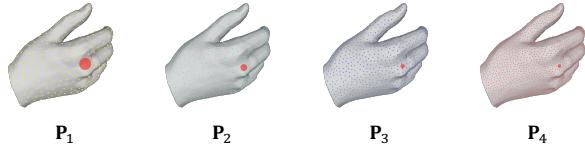


Figure B. Illustration of \mathbf{P}_k in different levels of resolutions. The light red area in the figure indicates the region used for spatial interpolation for a query point.

A.2. Network Structure

Architecture. For shape fitting, \mathcal{M}_{shape} consists of 4 layers with a hidden dimension of 128. For multi-resolution fields, \mathcal{M}_k contains 4 layers with a hidden dimension of 256, and \mathcal{M}_{fuse} contains 3 layers with a hidden size of 64.

Model Size. The trainable parameters of the model are 4.46M in total.

A.3. Used Data

InterHand2.6M [53]. For hand prior learning, we utilize ‘train/Capture0’, ‘train/Capture1’, ‘train/Capture2’, ‘train/Capture3’, ‘train/Capture5’, ‘train/Capture6’, ‘train/Capture7’, ‘train/Capture8’, ‘train/Capture9’, ‘train/Capture10’, ‘train/Capture11’, ‘train/Capture12’, ‘train/Capture13’, ‘train/Capture14’, ‘train/Capture15’, ‘train/Capture16’, ‘train/Capture20’, ‘train/Capture22’, ‘train/Capture23’, ‘train/Capture24’, and ‘train/Capture25’ for training, with one capture for an identity. Moreover, we use ‘0000_neutral_relaxed’, ‘0009_thumbtucknormal’, ‘0019_alligator_closed’, ‘0029_indextip’, ‘0039_fingerspreadrigid’, ‘0048_index_point’, and ‘0058_middlefinger’ for evaluation and other poses for training. For evaluation of one-shot avatars, we employ ‘test/Capture0/ROM03_RT_No_Occlusion’, following HandAvatar [9]. For each frame, we crop the hand region with annotated detection boxes as the ground truth, which is consistent with HandAvatar. Specifically, the box is firstly regulated as a square box with 1.3 times expansion. Then, the hand region is cropped and resized to 256×256 resolution. Unless otherwise stated, we all adopt the same cropping approach for experiments of all the used data.

HanCo [86]. For quantitative experiments on the HanCo dataset, we utilize sequence ‘0191’ with the camera above the hands (cam3, cam5, cam6, and cam7). We do not adopt cameras below the hands because the capturing environment exhibits uneven lighting conditions and inconsistent color calibration, which causes significantly inconsistent appearances of the hands for the images captured below. We utilize the MANO annotations and the provided hand

masks of this dataset for one-shot reconstruction.

MSCOCO [40]. To test OHTA’s performance for the challenging in-the-wild images, we take the whole-body version of MSCOCO [25] for experiments. We utilize the pose estimation results provided by InterWild [52] and generate the masks using SAM [33].

OneHand10K [73]. In addition to MSCOCO, we also provide in-the-wild results on the OneHand10K dataset, which is one of the largest monocular hand pose estimation datasets. Since OneHand10K does not provide labels for 3D pose estimation of the hand, we utilize DIR [64] for pose estimation and use the corresponding pose estimation results to generate hand masks.

Real-captured Data. We also capture images of hands with different identities and poses, ensuring a large demographic diversity among the subjects to show the robustness of OHTA. The hand pose estimation results and masks for those images are obtained by DIR [64].

A.4. Hand Prior Learning

The identity codes $\mathbf{z} \in \mathbb{R}^{21 \times 33}$ are learnable parameters initialized from a truncated normal distribution with standard deviation $\sigma = 0.02$, where 21 denotes that we utilize 21 subjects of InterHand2.6M [53] for training and 33 denotes the dimension of an identity’s code.

We follow the implementation of HandAvatar [9] to pre-train PairOF. If not stated otherwise, we adopt Adam optimizer [32] for optimization. The learning rate begins at $5e^{-4}$ and decreases with exponential decay. The training process has 300K steps with a batch size of 32.

For end-to-end prior learning, we use a patch strategy for training with a patch size 32×32 , following [9, 75]. The learning rate begins at $5e^{-4}$ and decreases with exponential decay. The complete prior learning process takes 300K steps with a batch size of 2. Since the masks from MANO are not well aligned with the hands in the image, using those masks for personalized shape fitting will result in underperforming learning of texture prior. Therefore, we adopt SAM [33] for mask prediction since it can obtain hand contour-aligned results. Specifically, we utilize the ViT-H version with 2D joint positions as the point prompts for predictions.

A.5. One-shot Reconstruction

Texture Inversion. For inversion, the input and rendered image resolutions are 256×256 . We initialize the identity code as zero vectors for optimization. The complete inversion process takes 50 steps with the learning rate of $1e^{-2}$. Since the fingernails are not related to color calibration and might dominate the inversion, we mask the fingernails for optimization. Specifically, we utilize the classified occupancy values of PairOF [9] to derive the masks of the fingernails.

Texture Fitting. For fitting, we use a patch strategy with a

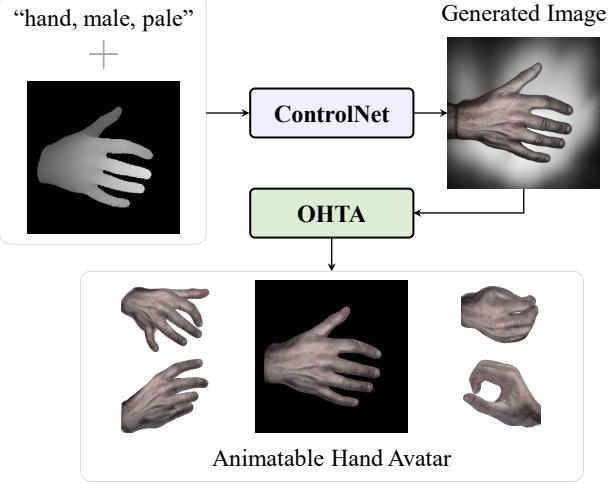


Figure C. Hand avatar creation with text prompts.

patch size 128×128 for optimization. The complete inversion process takes 100 steps with the learning rate of $1e^{-3}$. The N^r reference views for view regularization are generated by uniformly rotating the hands around the axis of the MCP joint of the middle finger and wrist with each rotation angle of $\frac{2\pi}{(N^r+1)}$. The hand pose of the reference views are set to flatting hand (canonical pose of MANO). Considering the fingernails may vary a lot from the inversion result to the fitting target, we make use of an additional loss for the fingernails using the fingernails mask derived from the PairOF.

A.6. Application

Text-to-avatar. The complete pipeline for text-to-avatar is illustrated in Fig. C. We use ControlNet 1.1 [80] with depth maps as inputs for hand image generation. The generation is guided by the depth map of the back of the hand and input text prompts, using the default parameters of the model. After generation, we utilize OHTA to reconstruct hand avatars from the hand images. The optimizing process follows the same procedure of one-shot reconstruction (Sec. A.5) except that using 200 steps for texture fitting to better capture complex details of the generated hands.

Editing. On account of our geometry based on the mesh scaffold, we can edit the geometry of avatars by changing the guided mesh scaffold. Since we utilize the MANO for guidance, we can modify the shape parameters β of the MANO to edit the geometry. For appearance editing, we can first render a target view of the hand avatar and draw desired content on the rendered image. Then, we can utilize OHTA for one-shot reconstruction. This process does not require texture inversion and only makes use of 100 steps’ texture fitting with the edited contents and the corresponding mask to update the edited parts.

Latent space manipulation. With multiple identities for

training, we obtain a continuous latent space. Therefore, we can conduct latent space manipulation, including latent space sampling and interpolation. For sampling, we randomly sample an identity code $\mathbf{z}' \in \mathbb{R}^{33}$ from a normal distribution. Then, we can use \mathbf{z}' to obtain the sampled avatar. For interpolation, we can take two identity codes \mathbf{z}_1 and \mathbf{z}_2 for combinations: $\mathbf{z}'' = t\mathbf{z}_1 + (1 - t)\mathbf{z}_2$, where $t \in [0, 1]$. With the interpolated identity code \mathbf{z}'' , we can obtain the hand avatar with the interpolated appearance.

B. More Experiments

Comparison with Handy. As shown in Fig. D, we compare our results with the in-the-wild results presented in the Handy [60] paper. The experimental results demonstrate that OHTA is capable of effectively modeling variations in skin tone and the details of the hand.

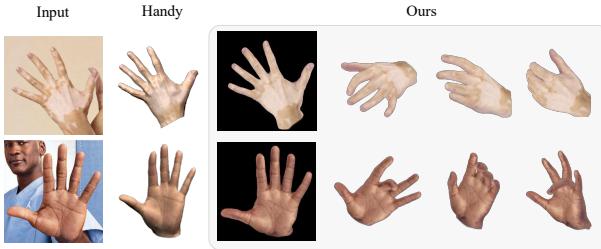


Figure D. Comparison with Handy on the results reported in the original paper.

Different Resolution. Using high-resolution encodings with dense points is able to model details of the texture, while low-resolution encodings with sparse points are capable of modeling the long-range dependencies of the texture for a more consistent appearance. Therefore, we conduct experiments to tune the best resolution combinations. Tab. A shows the performance of Hand Prior Network (HP-Net) with different resolutions for the albedo prior learning. We do not take resolutions larger than 4096 points since it will introduce too much computation. The results show that using 4 resolutions performs best. More resolutions (*e.g.*, 8 resolutions) degenerating the performance may be because 1) resolutions of too sparse points (*e.g.*, $\{32 \times 2^{k-1}\}_{k=1}^4$) are not able to encode detailed information is beneficial for the performance and 2) too many resolutions may lead to optimization difficulty. Therefore, we adopt 4 resolutions for our HPNet. The qualitative comparisons between using multi-resolution and single-resolution are shown in Fig. E. The results are tested with unseen poses. From those results, we can see that using multi-resolution is able to model more details of the hands and makes the learned overall appearance more consistent with the ground-truth.

Different Mask. As shown in Fig. F, using masks better aligned with the images is beneficial for personalized shape fitting during hand prior learning stage. When the hand ge-

Used Resolution	PSNR	LPIPS	SSIM
$\{4096\}$	27.11	12.96	0.890
$\{512, 1024\}$	27.18	12.91	0.890
$\{512, 4096\}$	27.46	12.80	0.895
$\{512 \times 2^{k-1}\}_{k=1}^4$	27.64	12.23	0.896
$\{32 \times 2^{k-1}\}_{k=1}^8$	27.32	12.28	0.894

Table A. Comparison of using different resolution combinations for the albedo field.



Figure E. Comparison between using multi-resolution and single-resolution. The red ellipse indicates those not well-captured details by single-resolution.

ometry is refined to be better aligned with the input images using the masks from SAM, the learned texture prior can better capture the details of the hand to improve the fidelity.

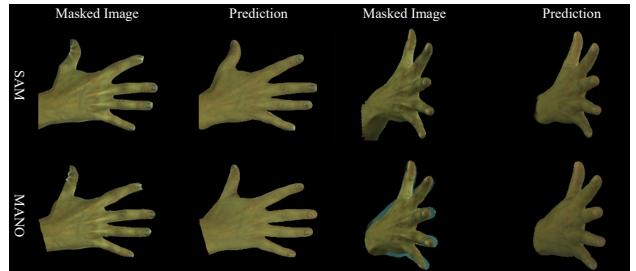


Figure F. Comparison between using different masks.

C. Discussion

C.1. Mesh-guided Representation

We adopt mesh-guided representation for several reasons. For the geometry, the implicit occupancy field learned with the mesh information is more robust for novel identities, as proved in [50]. For the texture, the implicit texture field learned on the mesh scaffolds can be transferred to other hand shapes, which is shown by our geometry editing capability. This nature is important for the one-shot reconstruction since the hands have a large variety in shape. Other

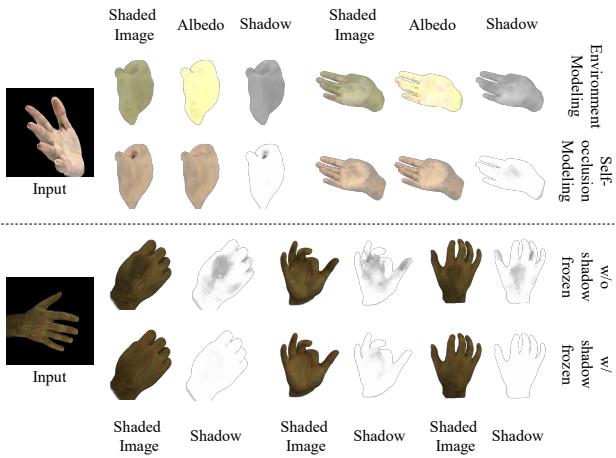


Figure G. Comparison between using different shadow approaches for one-shot reconstruction. The above shows the comparisons between environment modeling and our self-occlusion modeling. The below illustrates the differences between whether to freeze the shadow field for one-shot reconstruction.

texture representations (e.g., volume-based) [16, 55] have poor performance for novel hand shapes since they learn the neural hand representations in canonical space that have the fixed hand size. Moreover, this representation is more robust for one-shot fitting since it utilizes several neighboring anchors’ features for prediction to obtain relatively transition-smooth results.

C.2. Shadow Modeling

HPNet separates texture modeling into albedo and shadow modeling, following a foundational principle similar to that of HandAvatar [9]. However, our design diverges significantly from HandAvatar, particularly in its motivation and network structure.

For motivation, we aim to learn transferable shadow prior knowledge for one-shot hand avatar creations, while HandAvatar learns the illumination fields for a specific hand that are not transferable. To this end, our shadow branch is designed to be identity-shared and view-independent, which can learn transferable general self-occlusion effects that are shared across different identities. The reasons for this design are twofold. On one hand, it is non-trivial to learn the environmental lighting conditions based on a single image. Thus, we should rely on shadow prior learned from training data to generate plausible shadows for one-shot hand avatar creations. On the other hand, the identity-shared and view-independent shadow prior can be transferred to one-shot creation, while the identity-specific and view-conditioned prior learned by environmental lighting modeling cannot.

The specific shadow network structure of HPNet includes the identity-shared design, hand finger pose without global rotation as the pose condition, and the same multi-



Figure H. Part of the learned identities in the prior learning stage.

resolution fields as our albedo field. In comparison, HandAvatar has a dedicated design for an illumination field with positional encodings, directed soft occupancy, and a pose with global rotation as input, enabling it to capture the environmental lighting conditions for different hand poses.

As shown in Fig. G, our learned shadow prior can be effectively transferred to the novel identity to predict the identity-shared self-occlusion effects when there is no adequate knowledge of the real environmental lighting condition. In contrast, using identity-specific view-conditioned environment modeling like HandAvatar [9] fails in reconstructing the novel hand avatar with plausible shadows.

C.3. Difference between hand and other body parts

While recent studies have focused on one-shot human body or head avatars, they are not suitable for hand modeling due to inherent task differences. Created hand avatars aim to produce highly consistent animations for different poses and viewpoints, which is not similar to head avatars that concentrate less on the performance around the backside of the heads and are even not concerned about the animation like Preface [6]. Human avatars also require animation considerations, yet existing methodologies are inadequate for hand animations. In the case of explicit approaches employing generative models, such as DINAR [70], our experiments detailed in the main text for Handy [60] show that these methods fall short in generating high-fidelity hand avatars. Approaches employing diverse model priors like ELICIT [20] mainly utilize the symmetry of human body appearance for texture modeling. This kind of method is not appropriate for hands since modeling the hand appearance is mainly about modeling the details of the hands rather than only the hand skin tone. Those priors are not enough to model invisible hand details. Another type of generalizable NeRF approach, like SHERF [19], may also be inapplicable for one-shot hand avatar creation due to its inability to model high-quality unobserved parts of hands.

C.4. Discussion about OHTA and PhoneScan

PhoneScan [7] also proposes using prior models coupled with fine-tuning to achieve few-shot reconstructions. We provide a discussion about the similarities and differences between OHTA and PhoneScan.

Similarity: Both OHTA and PhoneScan 1) embed the priors to the model by training on large-scale data, and 2) include a fine-tuning process for adjusting the pre-trained model to fit the target images.

Difference: OHTA differs from PhoneScan in various aspects. 1) *motivation*: OHTA aims at addressing hand avatar creation from a **single** RGB image, while PhoneScan focuses on solving head avatar creation from RGB-D videos. Even with methods shared similarities, this essential difference leads to different design priorities (described below); 2) *optimization pipeline*: OHTA reconstructs with inversion and fitting, which highly relies on inversion from identity space for regularization of the fitting. As described in [7], PhoneScan has no identity space, predicts person-specific information (*i.e.* bias maps) with the pre-trained model, and fine-tunes the model with inputs. As mentioned in the ablation of [7], there exist artifacts for novel views when only fine-tuning with frontal images. In contrast, OHTA exhibits robust performance for one-shot creation.

3) *prior representation*: OHTA exploits identity codes and MLPs for geometry, albedo, and shadow prior learning, while PhoneScan embeds identity and expression prior knowledge in pre-trained CNNs.

C.5. Correlation between input and identity codes

In Fig. I, we show the identity codes distribution with different inputs, using t-SNE to embed the identity codes into 2D space. We found that similar inputs have closer distances in the identity space. For one-shot creation, identity codes 1) only relate to the hand albedo and 2) do not affect geometry, based on our hand representation design. Fig. 1 in the main text demonstrates that our identity space is continuous, allowing for interpolation between two hand identities while maintaining the hand geometry. Fig. R in supp. shows OHTA’s capability for shape editing while preserving the identity.

D. Qualitative Result

D.1. Prior Learning Result

Fig. H presents some of the learned identities in the prior learning stage. The images are rendered with different identity codes and the same hand pose. Those results demonstrate that our prior learning is able to capture the details of different identities, which facilitates the one-shot reconstruction.

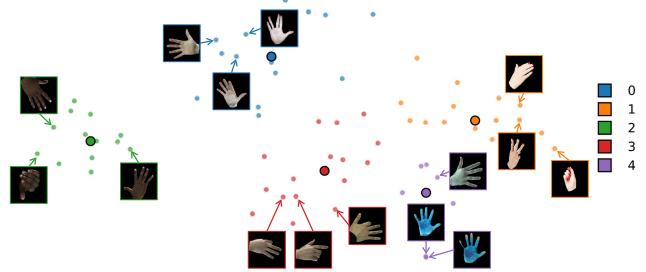


Figure I. t-SNE visualization of optimized identity codes from diverse images. Labels of clustering (colored points) are shown for visualization.

D.2. One-shot Result

To better show the robustness of OHTA, we show quantities of reconstructed hand avatars from OHTA on InterHand2.6M [53], HanCo [86], MSCOCO [25, 40], OneHand10K [73], and real-captured data. The input images of InterHand2.6M are all from the testing set that has no identity and sample overlap with the prior learning stage. As shown in Fig. K, our performance on InterHand2.6M is consistent across diverse hand poses and identities. We also take lots of samples varied a lot from the hands for prior learning (InterHand2.6M) to further demonstrate the one-shot capability of OHTA. Fig. L shows more visual results of OHTA on HanCo. OHTA is quite robust for different poses and viewpoints of the HanCo dataset. We present more results of OHTA on the MSCOCO dataset in Fig. M. To better validate the robustness of OHTA, we also take lots of samples of the OneHand10K dataset for experiments. The results are shown in Fig. N. Reconstructing hand avatars for MSCOCO and OneHand10K is challenging because 1) the pose estimations of the hands exhibit apparent deviations from the hands in the images and 2) the hands are with low resolutions. Despite facing these challenges, OHTA is still robust enough to create hand avatars with consistent animations. Fig. O presents more hand avatars reconstructed by OHTA for the real-captured images.

D.3. Application Result

We provide more results for text-to-avatar in Fig. P, appearance editing in Fig. Q, and shape editing in Fig. R. The animatable hand avatars from text prompts and appearance editing justify OHTA’s capability to capture the highly complex details for high-fidelity modeling. Moreover, the shape editing further validates that the mesh-guided design can fully transfer the texture prior to the novel hand shape, which is essential for the one-shot reconstruction.

D.4. Video Demo

We also provide additional qualitative results in the attached video.

E. Limitations and Failure Cases

In this section, we outline the limitations of our method to provide a more complete sense of the work’s scope.

OHTA relies on pose estimations for creating avatars. For some challenging cases as shown in Fig. J, e.g., ① challenging lighting conditions and ② skewed hand images, OHTA may fail because the estimated poses have significant errors or completely fail. However, if the pose estimation is reliable, OHTA has the robustness to produce results.

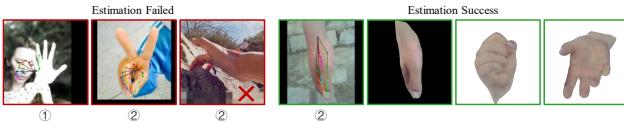


Figure J. Failure cases of the estimator (left). With successful estimations, OHTA creates hand avatars (right). The images come from the Onehand10K [73] dataset or the internet.

For 1) notably uneven lighting, and 2) some challenging poses with inaccurate estimations, OHTA may still fail. OHTA cannot resolve notably uneven lighting since it cannot detect the hand shadows on the input image. As shown in the first row of Fig. S, using images with severe shadows for one-shot reconstruction often results in undesirable shadows on avatars and unnatural transitions between observed and unobserved parts. OHTA relies on the estimated pose results for geometry modeling, which forms the basis for texture modeling. Therefore, poor estimation results lead to inferior texture modeling, especially for challenging poses. Even though OHTA is robust to inaccurate estimations to some extent, as shown in the results for in-the-wild images, due to the learned hand priors, it is still incapable of addressing highly inaccurate estimations for some challenging poses. The second row of Fig. S illustrates the texture misalignment and artifacts caused by those challenging poses with inaccurate estimations.

Another limitation relates to the number of fingers. Since the estimator and OHTA both utilize a hand parametric model (e.g. MANO) with only five fingers. It is hard to resolve hands with more or less than five fingers. We believe that exploring the use of non-parametric representations could help address this challenge.



Figure K. **Qualitative results on InterHand2.6M [53].** For each example, we show from left to right: (a) the target image, (b) the fitted avatar rendered to the input view, and (c) the results of the hand avatar rendered in novel poses.

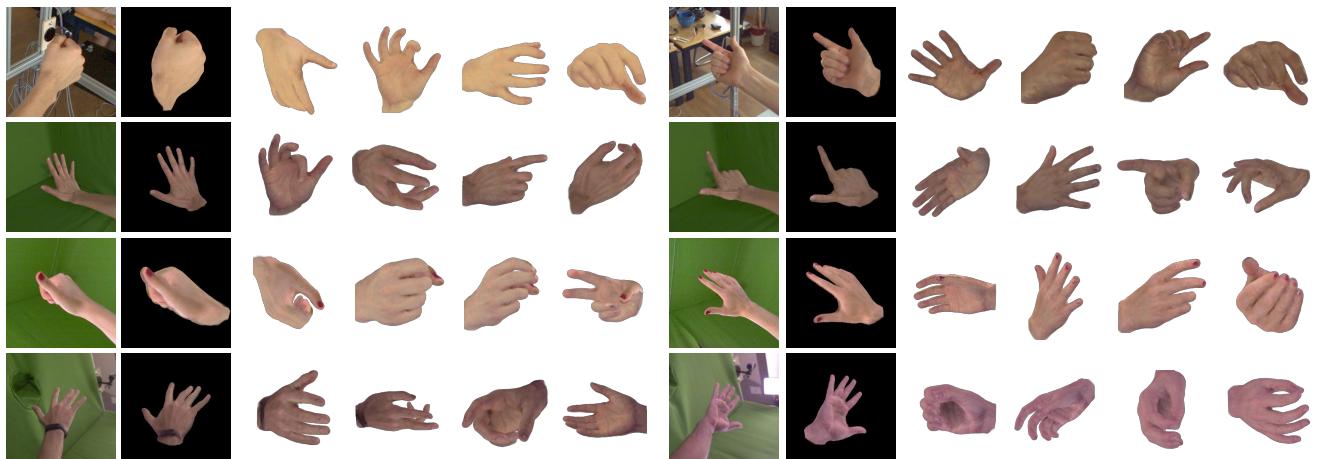


Figure L. **Qualitative results on HanCo [86].** For each example, we show from left to right: (a) the target image, (b) the fitted avatar rendered to the input view, and (c) the results of the hand avatar rendered in novel poses.



Figure M. **In-the-wild results on MSCOCO [40].** For each example, we show from left to right: (a) the target image, (b) the fitted avatar rendered to the input view, and (c) the results of the hand avatar rendered in novel poses.



Figure N. **In-the-wild results on OneHand10K [73].** For each example, we show from left to right: (a) the target image, (b) the fitted avatar rendered to the input view, and (c) the results of the hand avatar rendered in novel poses.



Figure O. **In-the-wild results on the real-captured data.** For each example, we show from left to right: (a) the target image, (b) the fitted avatar rendered to the input view, and (c) the results of the hand avatar rendered in novel poses.



Figure P. Text-to-avatar results.



Figure Q. Appearance editing results upon one-shot avatars.



Figure R. Shape editing results upon one-shot avatars.



Figure S. Failure cases due to notably uneven lighting (in the first row) or highly inaccurate pose estimation results (in the second row).

References

- [1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. [2](#)
- [2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2293–2303, 2019. [2](#)
- [3] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5461–5470, 2021. [3](#)
- [4] Ziqian Bai, Feitong Tan, Zeng Huang, Kripasindhu Sarkar, Danhang Tang, Di Qiu, Abhimitra Meka, Ruofei Du, Mingsong Dou, Sergio Orts-Escalano, et al. Learning personalized high quality volumetric head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16890–16900, 2023. [3](#)
- [5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *The European Conference on Computer Vision*, pages 561–578, 2016. [2](#)
- [6] Marcel C Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escalano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, et al. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3402–3413, 2023. [2, 3, 12](#)
- [7] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shouo-I Yu, et al. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. [3, 4, 13](#)
- [8] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. [3](#)
- [9] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8683–8693, 2023. [1, 2, 3, 4, 5, 6, 7, 9, 10, 12](#)
- [10] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10451–10460, 2021. [1](#)
- [11] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20533–20543, 2022. [1, 2, 5, 6](#)
- [12] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *The European Conference on Computer Vision*, pages 612–628, 2020. [3](#)
- [13] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. Dart: Articulated hand model with diverse accessories and rich textures. *Advances in Neural Information Processing Systems*, 35:37055–37067, 2022. [1](#)
- [14] Daiheng Gao, Xindi Zhang, Xingyu Chen, Andong Tan, Bang Zhang, Pan Pan, and Ping Tan. Cyclehand: Increasing 3d pose estimation ability on in-the-wild monocular image through cyclic flow. In *Proceedings of the ACM International Conference on Multimedia*, pages 2452–2463, 2022. [1](#)
- [15] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. [3](#)
- [16] Zhiyang Guo, Wengang Zhou, Min Wang, Li Li, and Houqiang Li. Handnerf: Neural radiance fields for animatable interacting hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21078–21087, 2023. [1, 2, 12](#)
- [17] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33:9276–9287, 2020. [2](#)
- [18] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11046–11056, 2021. [2](#)
- [19] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. [3, 12](#)
- [20] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. [3, 12](#)
- [21] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. [2](#)
- [22] Zhisheng Huang, Yujin Chen, Di Kang, Jinlu Zhang, and Zhigang Tu. Phrit: Parametric hand representation with implicit template. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14974–14984, 2023. [3](#)

- [23] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Sel-frecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 2, 6, 7
- [24] Yuheng Jiang, Kaixin Yao, Zhuo Su, Zhehao Shen, Haimin Luo, and Lan Xu. Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 595–605, 2023. 2
- [25] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *The European Conference on Computer Vision*, pages 196–214, 2020. 8, 10, 13
- [26] Kyungmin Jo, Gyumin Shim, Sanghun Jung, Soyoung Yang, and Jaegul Choo. Cg-nerf: Conditional generative neural radiance fields for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 724–733, 2023. 3
- [27] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 2
- [28] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2
- [29] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2
- [30] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *International Conference on 3D Vision*, pages 11–21, 2021. 3
- [31] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. Harp: Personalized hand reconstruction from a monocular rgb video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12802–12813, 2023. 1, 2, 6, 7
- [32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 10
- [33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 10
- [34] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2
- [35] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019.
- [36] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *International Conference on 3D Vision*, pages 643–653, 2019. 2
- [37] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36(6):194–1, 2017. 3
- [38] Yuwei Li, Minye Wu, Yuyao Zhang, Lan Xu, and Jingyi Yu. Piano: A parametric hand bone model from magnetic resonance imaging. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 816–822, 2021. 2
- [39] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuxin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 1, 2
- [40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *The European Conference on Computer Vision*, pages 740–755, 2014. 8, 10, 13, 16
- [41] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261*, 2023. 3
- [42] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 2023. 3, 6
- [43] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 3
- [44] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 4
- [45] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 3
- [46] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2, 3
- [47] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 2
- [48] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks:

- Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 3
- [49] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10461–10471, 2021. 3
- [50] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. Coap: Compositional articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13201–13210, 2022. 3, 4, 11
- [51] 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. In *The European Conference on Computer Vision*, pages 405–421, 2020. 2, 5
- [52] Gyeongsik Moon. Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17028–17037, 2023. 10
- [53] 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 *The European Conference on Computer Vision*, pages 548–564, 2020. 5, 6, 9, 10, 13, 15
- [54] Jiteng Mu, Shen Sang, Nuno Vasconcelos, and Xiaolong Wang. Actorsnerf: Animatable few-shot human rendering with generalizable nerfs. pages 18391–18401, 2023. 3
- [55] Akshay Mundra, Jiayi Wang, Marc Habermann, Christian Theobalt, Mohamed Elgharib, et al. Livehand: Real-time and photorealistic neural hand rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 5, 6, 7, 12
- [56] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. Star: Sparse trained articulated human body regressor. In *The European Conference on Computer Vision*, pages 598–613, 2020. 2
- [57] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2, 3
- [58] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2
- [59] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 3
- [60] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4670–4680, 2023. 1, 2, 6, 7, 11, 12
- [61] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *The 12th International Conference on Learning Representations*, 2024. 3
- [62] 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 *The European Conference on Computer Vision*, pages 54–71. Springer, 2020. 1, 2
- [63] Wentian Qu, Zhaopeng Cui, Yinda Zhang, Chenyu Meng, Cuixia Ma, Xiaoming Deng, and Hongan Wang. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15100–15111, 2023. 1, 2
- [64] Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8014–8025, 2023. 5, 10
- [65] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017. 2, 3, 4
- [66] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morigama, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2
- [67] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2
- [68] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 3
- [69] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021. 2
- [70] David Svitov, Dmitrii Gudkov, Renat Bashirov, and Victor Lempitsky. Dinar: Diffusion inpainting of neural textures for one-shot human avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7062–7072, 2023. 2, 12
- [71] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proceedings of*

- the IEEE/CVF International Conference on Computer Vision*, pages 11708–11718, 2021. 3
- [72] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. *Advances in Neural Information Processing Systems*, 34:2810–2822, 2021. 3
- [73] Yangang Wang, Cong Peng, and Yebin Liu. Mask-pose cascaded cnn for 2d hand pose estimation from single color image. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11):3258–3268, 2018. 10, 13, 14, 17
- [74] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6
- [75] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira 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*, pages 16210–16220, 2022. 2, 6, 7, 10
- [76] Tong Wu, Jia-Mu Sun, Yu-Kun Lai, and Lin Gao. De-nerf: Decoupled neural radiance fields for view-consistent appearance editing and high-frequency environmental relighting. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3
- [77] Hongyi Xu, Thiendo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 2
- [78] Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *The European Conference on Computer Vision*, pages 597–614, 2022. 3
- [79] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 3
- [80] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 8, 10
- [81] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5, 6
- [82] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humanerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 2
- [83] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021. 2
- [84] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022. 3
- [85] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4491–4500, 2019. 2
- [86] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *DAGM German Conference on Pattern Recognition*, pages 250–264. Springer, 2021. 6, 9, 13, 15