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 $k \in \{1, ..., K\}$ 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:

$$\mathbf{Q}_k = \text{interp}(\mathbf{q}, \mathbf{E}_k). \tag{1}$$

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).

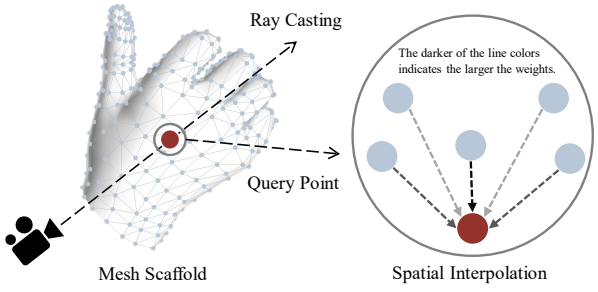

Figure A. Illustration of spatial interpolation for a level of resolution.

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.

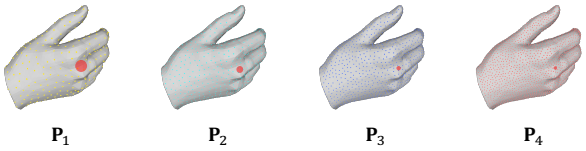

Figure B. Illustration of $\mathbf{P}_k$ in different level of resolutions. The light red area in the figure indicates the region used for spatial interpolation for a query point.

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 [13]. 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 [3]. 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 [21]. 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 [10]. To test OHTA's performance for the challenging in-the-wild images, we take the whole-body version of MSCOCO [7] for experiments. We utilize the pose estimation results provided by InterWild [12] and generate the masks using SAM [9].

OneHand10K [18]. 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 [16] 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 [16].

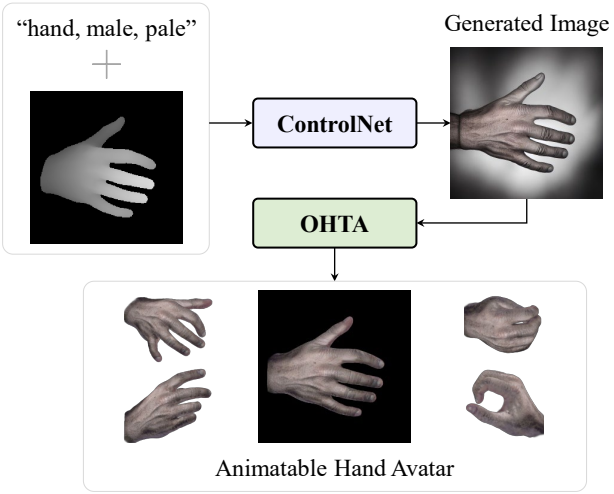

Figure C. Hand avatar creation with text prompts.

A.4. Hand Prior Learning

The identity codes $\mathbf{z} \in \mathbb{R}^{21\times33}$ 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 [13] for training and 33 denotes the dimension of an identity's code.

We follow the implementation of HandAvatar [3] to pretrain PairOF. If not stated otherwise, we adopt Adam optimizer [8] 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 [3, 19]. 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 [9] 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 [3] to derive the masks of the fingernails.

Texture Fitting. For fitting, we use a patch strategy with a 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 [20] 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 $\boldsymbol{\beta}$ 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 [15] paper. The experimental results demonstrate that OHTA is capable of effectively modeling variations in skin tone and the details of the hand.

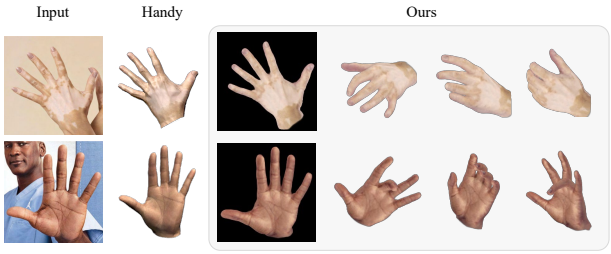

Figure D. Comparison with Handy on the results reported in the original paper.

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 (HPNet) 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 geometry 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.

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.

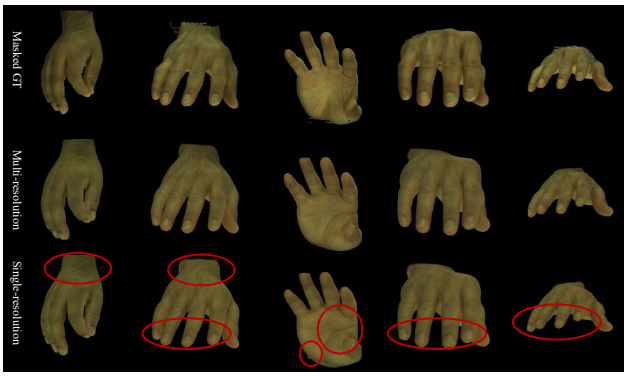

Figure E. Comparison between using multi-resolution and single-resolution. The red ellipse indicates those not well-captured details by single-resolution.

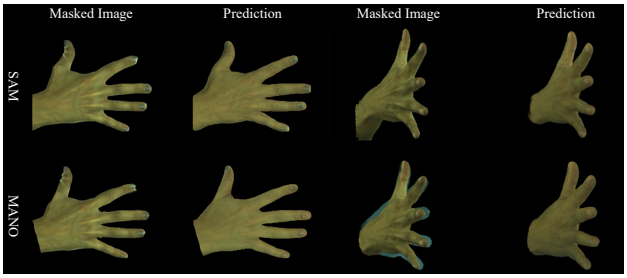

Figure F. Comparison between using different masks.

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 [11]. For the texture, the implicit texture field learned on the mesh scaffolds can be transferred to other

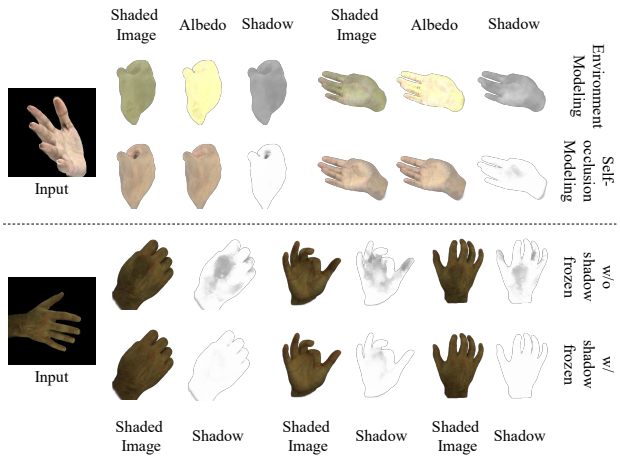

Figure G. Comparison between using different shadow approaches for one-shot reconstruction. The above shows the comparisons between environment modeling and our self-occlusion modeling. The below illustrates the differences between whether to freeze the shadow field for one-shot reconstruction.

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 texture representations (e.g., volume-based) [4, 14] 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 [3]. 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 transferrable 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.

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

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-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 [3] 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 [1]. 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 [17], our experiments detailed in the main text for Handy [15] show that these methods fall short in generating high-fidelity hand avatars. Approaches employing diverse model priors like ELICIT [6] 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 [5], 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 [2] 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 [2], 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 [2], 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.

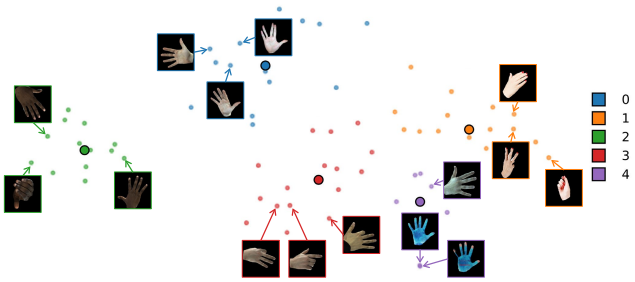

Figure I. t-SNE visualization of optimized identity codes from diverse images. Labels of clustering (colored points) are shown for visualization.

D.2. One-shot Result

To better show the robustness of OHTA, we show quantities of reconstructed hand avatars from OHTA on InterHand2.6M [13], HanCo [21], MSCOCO [7, 10], OneHand10K [18], 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.

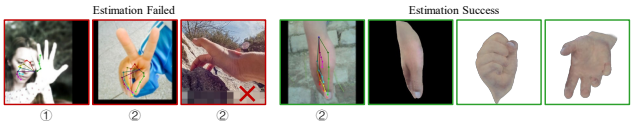

Figure J. Failure cases of the estimator (left). With successful estimations, OHTA creates hand avatars (right). The images come from the Onehand10K [18] dataset or the internet.

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 [13].** 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.

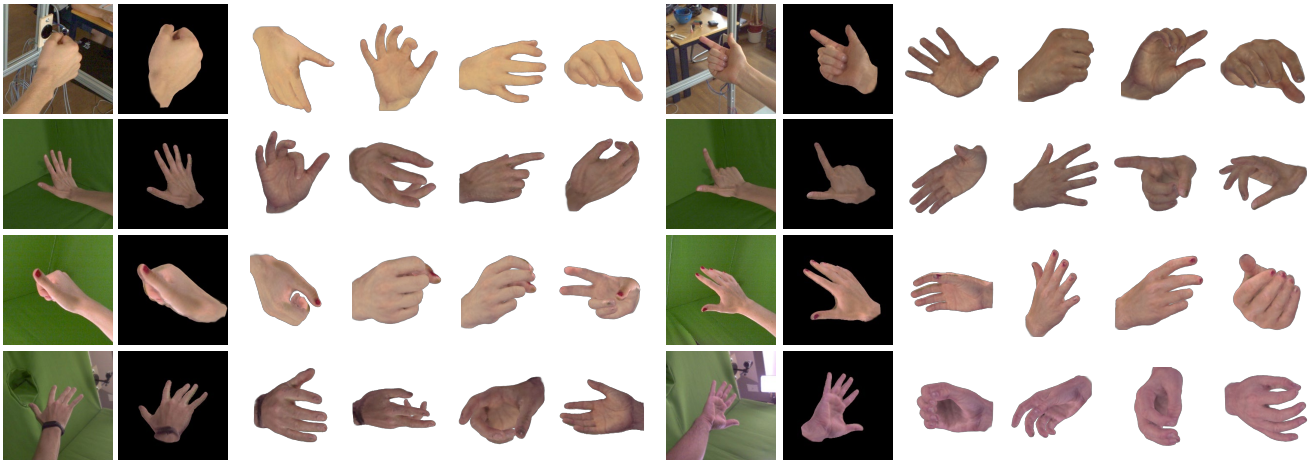

Figure L. **Qualitative results on HanCo [21].** For each example, we show from left to right: (a) the target image, (b) the fitted avatar rendered to the input view, and (c) the results of the hand avatar rendered in novel poses.

Figure M. **In-the-wild results on MSCOCO [10].** 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 [18].** 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.

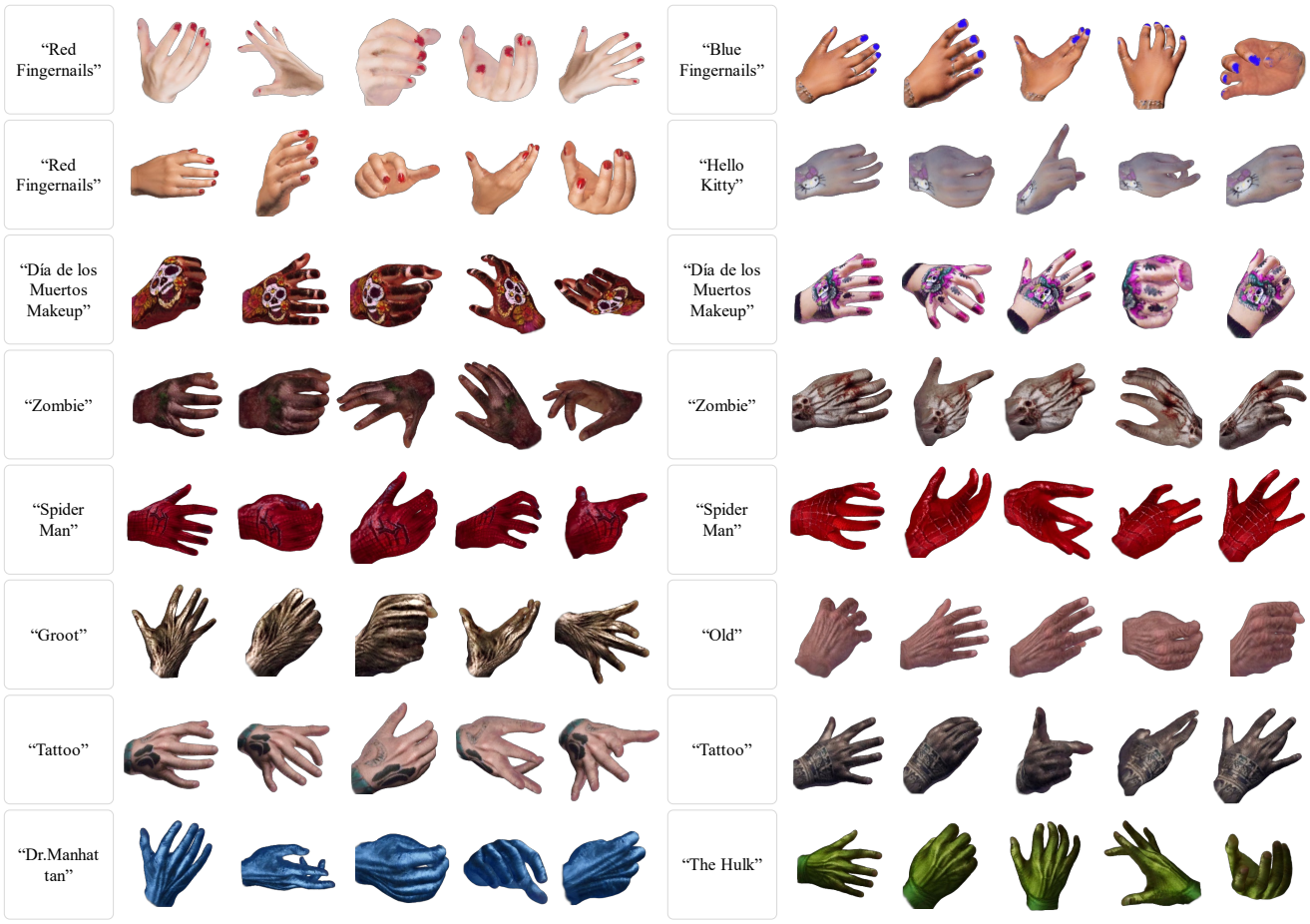

Figure P. Text-to-avatar results.

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] Marcel C Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, 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. 4

[2] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 5

[3] 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, 4

[4] 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. 4

[5] 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. 4

[6] 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. 4

[7] 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. 2, 5

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[9] 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. 2

[10] 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. 2, 5, 8

[11] 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

[12] 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. 2

[13] 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. 1, 2, 5, 7

[14] 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. 4

[15] 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. 3, 4

[16] 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. 2

[17] 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. 4

[18] 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. 2, 5, 6, 9

[19] 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

[20] 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. 2

[21] 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. 1, 5, 7