

# HiFiHR: Enhancing 3D Hand Reconstruction from a Single Image via High-Fidelity Texture

Jiayin Zhu, Zhuoran Zhao, Linlin Yang, and Angela Yao

National University of Singapore

{zhujiayin,zhuoran.zhao}@u.nus.edu, {yangll,ayao}@comp.nus.edu.sg

**Abstract.** We present HiFiHR, a high-fidelity hand reconstruction approach that utilizes render-and-compare in the learning-based framework from a single image, capable of generating visually plausible and accurate 3D hand meshes while recovering realistic textures. Our method achieves superior texture reconstruction by employing a parametric hand model with predefined texture assets, and by establishing a texture reconstruction consistency between the rendered and input images during training. Moreover, based on pretraining the network on an annotated dataset, we apply varying degrees of supervision using our pipeline, *i.e.*, self-supervision, weak supervision, and full supervision, and discuss the various levels of contributions of the learned high-fidelity textures in enhancing hand pose and shape estimation. Experimental results on public benchmarks including FreiHAND and HO-3D demonstrate that our method outperforms the state-of-the-art hand reconstruction methods in texture reconstruction quality while maintaining comparable accuracy in pose and shape estimation. Our code is available at <https://github.com/viridityzhu/HiFiHR>.

**Keywords:** 3D Hand Reconstruction · 3D from Single Images.

## 1 Introduction

With the development of VR/AR, photo-realistic 3D reconstruction has gained raising attention, and significant progress has been made in the human body and face [13,35,36,31,12]. Hands, as the main medium through which people interact with the world, are also essential to be accurately reconstructed.

Existing works [5,1,20,24,14] leverage differentiable rendering [3] and a parametric hand model, such as MANO [27], to reconstruct 3D hand meshes from images. These methods excel in achieving precise pose estimation while also demonstrating high efficiency during inference. Another recent line of work adapts NeRF-based [31] 3D reconstruction by modeling geometry and texture properties together from ray queries, resulting in high-quality textures [6,4,21]. However, these methods have limitations: (1) Render-based approaches often neglect texture reconstruction or achieve limited accuracy using simplistic representations. (2) NeRF-based methods require videos or multi-view images as input, and suffer from high computational complexity and limited generalization.

In practical applications, obtaining multi-view images can be time-consuming or unfeasible, leaving us with scenarios where only a single image is available. In such cases, render-based methods remain the most viable option. However, monocular images of hands commonly suffer from severe occlusion and depth ambiguity, posing challenges for reconstructing plausible hand structures. This motivates the need for high-fidelity texture reconstruction, as it has the potential to facilitate accurate estimation of hand pose and shape. In contrast, NIMBLE [18], an anatomy-based hand model, embeds numerous texture assets and leverages physical-based rendering [23] for a high-fidelity texture representation. Moreover, it physically constrains the texture by the relative motion between muscles, bones, and skins, which leads to reliable estimation results. Leveraging this novel texture representation holds great potential for enhancing hand reconstruction quality.

Building upon these challenges, our work focuses on achieving high-fidelity texture reconstruction in single-image scenarios. Leveraging the advantage of NIMBLE [18] in a rendering pipeline, we build a HiFiHR (High-fidelity hand reconstruction) model which is able to predict 3D hand pose, shape, texture, and lighting from single input images (Fig. 1).

Furthermore, previous works have attempted to enhance pose estimation through texture reconstruction. However, these efforts have primarily concentrated on single supervision settings, such as self-supervision by S<sup>2</sup>HAND [5] and weak supervision by SMHR [26]. Additionally, the limited quality of the reconstructed textures in these approaches hinders their performance. As a result, no significant conclusions have been drawn, leaving room for further exploration. In our work, we comprehensively investigate this question across various levels of supervision, aiming to ascertain the extent to which high-fidelity texture reconstruction consistency can aid the learning of pose and shape.

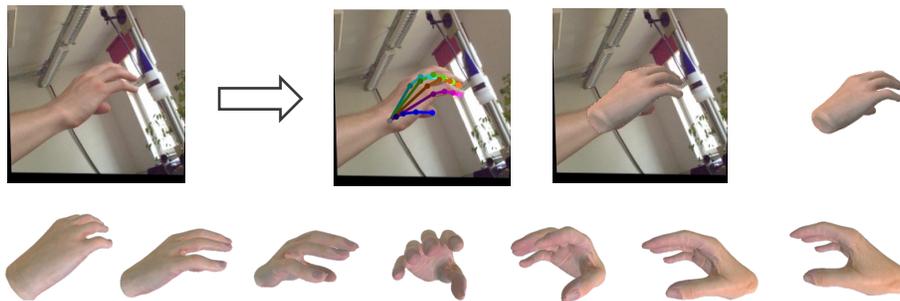
The main contributions of this work can be concluded as follows.

1. By leveraging the advantages of model-based methods, our approach produces high-fidelity and consistent hand textures from a single input image, resulting in a more realistic representation of the reconstructed 3D hands.
2. We investigate the impact of hand texture reconstruction on pose and shape estimation under varying levels of supervision. Our findings reveal that high-fidelity texture consistency aids pose and shape learning in self-supervision, but introduces noise in weak supervision. And the effect is minimal in full supervision with stronger constraints from 3D labels.
3. Quantitative and qualitative experiments on two single-hand reconstruction benchmarks, *i.e.*, FreiHAND and HO-3D, verify the effectiveness of our approach in both pose and shape accuracy and texture reconstruction quality.

## 2 Related Work

### 2.1 Parametric Hand Models

Given parameters on hand pose and shape, parametric hand models infer a hand mesh with vertices and faces. MANO [27] is a widely used hand model, which



**Fig. 1.** Our method reconstructs a realistic 3D hand from a monocular image leveraging the NIMBLE [18] hand model and the proposed texture reconstruction consistency. It synthesizes lifelike hands from just a single view, delivering plausible representations from every perspective.

follows SMPL [19] to adopt linear blend skinning to deform a mesh based on a kinematic skeleton. However, MANO does not capture a texture space. One follow-up work, HTML [24], captures hand texture from two image sequences and fits a MANO template to extract the texture from the scanned mesh. Furthermore, an anatomy-based hand model, NIMBLE [18], creates a better hand texture model via PCA and pre-defined texture assets. The photorealistic texture is represented with diffuse, normal, and specular maps, and a more reliable hand pose can be achieved by enforcing inner bones and muscles to match anatomic and kinematic rules.

## 2.2 Hand Texture Reconstruction

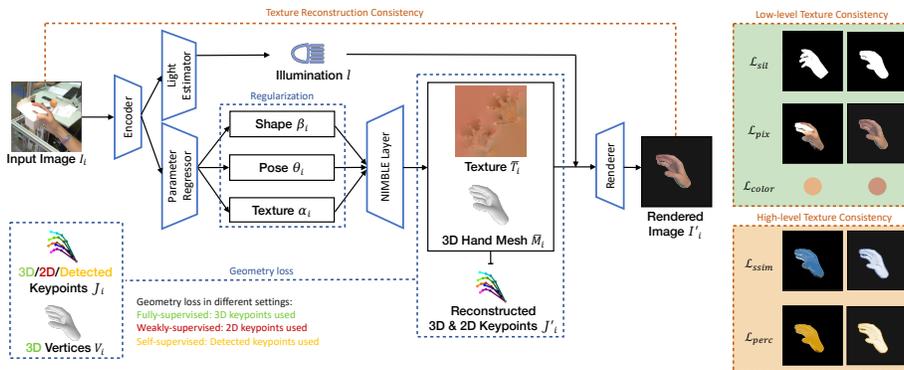
Boukhayma *et al.* [1] introduce the first end-to-end deep learning-based method that predicts both 3D hand shape and pose from RGB images in the wild. It leverages MANO as a pre-computed hand model and a re-projection module. Similarly, S<sup>2</sup>HAND [5] and SMHR [26] achieve 3D hand reconstruction with texture, also adopting MANO as a pre-computed hand model. However, these works achieve limited texture reconstruction because of the simple texture representation, *i.e.*, RGB values of each vertex in the mesh. DeepHandMesh [20] mainly focuses on shape reconstruction, while also providing a method to unwrap multiview RGB images to a high-quality  $1024 \times 1024$  texture map. Moreover, LiveHand [21] and HandAvatar [4] are two recent NeRF-based works that also achieve photo-realistic texture, but they are not suitable in scenarios when there is only one image available.

Differently, this work aims to achieve high-fidelity 3D hand texture reconstruction from a single image. Based on NIMBLE, a 3D hand mesh with high-resolution texture UV maps can be estimated from the 2D image. By employing a differentiable renderer, weak supervision on hand texture is achieved through consistency between the rendered 2D image and the input image. Additionally,

the pipeline incorporates pre-computed texture assets in NIMBLE, enabling the generation of plausible texture estimations for unseen or occluded parts of the hand, even when the input is derived from a monocular view. The iterative render-and-compare loop further ensures accurate alignment of the reconstructed 3D hand with the 2D image.

## 3 Methods

### 3.1 Overview



**Fig. 2.** Overview of our 3D hand reconstruction pipeline. Given one single-view image, our method generates a plausible and high-fidelity 3D textured hand mesh. The method is supervised by geometry loss, texture reconstruction consistency loss, and regularization. Geometry loss adopts different settings under different kinds of supervision. On the right are details of the texture reconstruction consistency, which consists of a low-level consistency and a high-level consistency.

Our end-to-end pipeline is able to reconstruct 3D hand given a single RGB image, as demonstrated in Fig. 2. Given the input image  $I_i$ , where  $i \in [1, N]$  and  $N$  is the number of samples in the dataset, we adopt an encoder  $\mathcal{E}$ , a NIMBLE [18] hand layer  $\mathcal{N}$ , and a differentiable renderer  $\mathcal{R}$ , to obtain a textured 3D hand mesh  $M_i$  and a rendered image  $I'_i$ .

### 3.2 Model Structure

**Parameter Encoder.** An encoder  $\mathcal{E}$  is trained to estimate the parameters including shape  $\beta_i$ , pose  $\theta_i$ , texture  $\alpha_i$ , and illumination  $l_i$ :

$$\{\beta_i, \theta_i, \alpha_i, l_i\} = \mathcal{E}(I_i). \quad (1)$$

The encoder consists of a visual feature extractor, *e.g.*, Efficientnet [32], and independent MLP layers for each parameter.  $\beta_i$ ,  $\theta_i$ , and  $\alpha_i$  are NIMBLE parameters controlling hand shape, pose, and appearance. They are vectors of length 20, 30, and 10, respectively, following the default setting of NIMBLE.

Disentangling lighting from texture is a crucial factor for obtaining plausible hand texture. This is because the feasible texture space of NIMBLE is limited to natural skin tones, while images in the wild often have more complex lighting conditions such as colorful lights that NIMBLE cannot fit. Therefore, we also estimate illumination  $l_i$  from the input image, which is a vector of length 6, defining directional lighting from a direction  $(x_i, y_i, z_i)$  with color  $(r_i, g_i, b_i)$ :

$$l_i = (x_i, y_i, z_i, r_i, g_i, b_i). \quad (2)$$

**NIMBLE Layer.** NIMBLE takes the parameters estimated by the encoder  $\mathcal{E}$  as input and generates a realistic 3D hand with bone, muscle, and skin geometry, as well as a photo-realistic appearance:

$$\{\bar{M}_i, \bar{T}_i\} = \mathcal{N}(\beta_i, \theta_i, \alpha_i), \quad (3)$$

where  $\bar{M}_i = \{M_i^{skin}, M_i^{bone}, M_i^{muscle}\}$ , which consists of a skin mesh, a bone mesh, and a muscle mesh. And  $\bar{T}_i = \{T_i^{diff}, T_i^{spec}, T_i^{norm}\}$ , which consists of a diffuse map, a specular map, and a normal map.

The skin mesh is physically conditioned by the bone mesh and muscle mesh to obtain plausible hand skin, which contains 5990 vertices and 9984 faces. 25 joints are obtained using a regressor from the skin mesh. However, commonly used datasets only provide annotations for the 778 vertices and 21 joints defined by MANO, and the 25 NIMBLE joints are not compatible. Thus, an additional linear layer is utilized to regress the 778 MANO vertices from the skin mesh, denoted as  $M_i$ . Additionally, we adopt the MANO regressor to obtain 21 MANO joints, denoted as  $J_i^{3D}$ , from  $M_i$ .

**Differentiable Render.** The generated textured mesh is then rendered back into the image domain as  $I'_i$  by a differentiable neural renderer PyTorch3D [25] utilizing the illumination  $l_i$  [3]:

$$I'_i = \mathcal{R}(M_i, T_i^{diff}, l_i, k_i), \quad (4)$$

where  $k_i$  is the ground truth camera intrinsic parameters. The 2D keypoints  $J'_i$  are also re-projected from 3D joints  $J_i^{3D}$  given  $k_i$ . Note that because the PyTorch3D renderer [25] does not support physical-based rendering yet, we only use  $T_i^{diff}$  as the texture UV mapping.

### 3.3 Training Objective

We train the pipeline using single-view hand images. The optimization objective is to minimize the difference between the input and the rendered hand images,

with different levels of supervision on joints and vertices. The overall training loss  $\mathcal{L}$  consists of three parts including geometry loss  $\mathcal{L}_{geo}$ , texture reconstruction consistency loss  $\mathcal{L}_{tex}$ , and regularization  $\mathcal{L}_{regu}$ :

$$\mathcal{L} = w_{geo}\mathcal{L}_{geo} + w_{tex}\mathcal{L}_{tex} + w_{regu}\mathcal{L}_{regu}, \quad (5)$$

where  $w_{geo}$ ,  $w_{tex}$ , and  $w_{regu}$  are weighting factors for each loss term.

**Geometry Loss.** The geometry of the mesh is constrained with the following terms:

$$\mathcal{L}_{geo}^{3D} = w_{jnt}\mathcal{L}_{jnt} + w_{vert}\mathcal{L}_{vert} + w_{direc}\mathcal{L}_{direc} + w_{len}\mathcal{L}_{len}, \quad (6)$$

$$\mathcal{L}_{geo}^{2D} = w_{jnt}\mathcal{L}_{jnt}^{2D} + w_{direc}\mathcal{L}_{direc}^{2D}. \quad (7)$$

$w_{jnt}$ ,  $w_{vert}$ ,  $w_{direc}$ , and  $w_{len}$  are weighting factors for each loss term. In the full 3D supervision scheme, we train the pipeline using  $\mathcal{L}_{geo}^{3D}$  where all the loss terms are in 3D space, *e.g.*, the 3D joint loss  $\mathcal{L}_{jnt}$ . While in the weak 2D supervision scheme, we only employ  $\mathcal{L}_{geo}^{2D}$  where all the loss terms are in 2D space after projection.  $\mathcal{L}_{jnt}$  and  $\mathcal{L}_{vert}$  penalize the  $l_1$ -distance between the predicted and ground-truth joint and vertex positions, respectively.  $\mathcal{L}_{direc}$  is the loss on bone directions, where a bone is defined as the vector between two adjacent joints.  $\mathcal{L}_{len}$  penalizes the difference in edge lengths of all faces. Furthermore, in the self-supervision scheme, we only utilize  $\mathcal{L}_{jnt}^{2D}$  and  $\mathcal{L}_{direc}^{2D}$  with detected 2D joints obtained through OpenPose [2] at an offline stage, along with corresponding bone directions. The confidence score provided by OpenPose is integrated into  $\mathcal{L}_{jnt}^{2D}$  and  $\mathcal{L}_{direc}^{2D}$ , which shows the probability of joints being detected correctly. Please refer to the supplementary for more information on the losses.

**Texture Reconstruction Consistency.** The texture reconstruction consistency is designed to ensure the consistency between the input and rendered images, and is formulated as:

$$\mathcal{L}_{tex} = \underbrace{w_{pix}\mathcal{L}_{pix} + w_{color}\mathcal{L}_{color} + w_{sil}\mathcal{L}_{sil}}_{\mathcal{L}_{low-level}} + \underbrace{w_{ssim}\mathcal{L}_{ssim} + w_{perc}\mathcal{L}_{perc}}_{\mathcal{L}_{high-level}}. \quad (8)$$

$\mathcal{L}_{pix}$  is a per-pixel  $l_1$  loss between the input and rendered images,  $\mathcal{L}_{color}$  encourages the mean RGB colors to be close,  $\mathcal{L}_{sil}$  penalizes the difference between silhouettes,  $\mathcal{L}_{ssim}$  measures the structural similarity between the two images [33], and  $\mathcal{L}_{perc}$  is the LPIPS loss [17,37] that captures the perceptual difference between the two images by comparing the features extracted using the AlexNet model [16]. We use a hand mask to extract the foreground hand of the input image to make a comparison with the rendered hand. In the full supervision and weak supervision schemes, we use the ground-truth hand mask. In the self-supervision scheme, we use the predicted hand mask.  $w_{pix}$ ,  $w_{color}$ ,  $w_{sil}$ ,  $w_{ssim}$ , and  $w_{perc}$  are weighting factors for each loss term.

Among these texture reconstruction consistency terms, the low-level texture consistency  $\mathcal{L}_{low-level}$  consists of  $\mathcal{L}_{pix}$ ,  $\mathcal{L}_{color}$ , and  $\mathcal{L}_{sil}$ , which operates at the low level of pixel values and results in sensitivity to geometry misalignment. To mitigate this issue, we also include the high-level texture consistency  $\mathcal{L}_{high-level}$ , consisting of  $\mathcal{L}_{ssim}$  and  $\mathcal{L}_{perc}$ , to enable abstract comparisons of the textures and align more closely with human perceptual results.

**Regularization.** We also add  $l_2$ -regularizers on the magnitude of the shape, pose, and texture parameters to ensure the results are plausible, where  $w_\beta$ ,  $w_\theta$ , and  $w_\alpha$  are weighting factors for each loss term:

$$\mathcal{L}_{regu} = w_\beta \|\beta\|_2 + w_\theta \|\theta\|_2 + w_\alpha \|\alpha\|_2. \quad (9)$$

## 4 Experiment

### 4.1 Implementation Details

The proposed method is implemented with PyTorch [22]. We use EfficientNet [32] pretrained on the ImageNet dataset as our feature extractor. The input to our model is a  $224 \times 224$  RGB image. In our training process, the batch size is set to 50. The initial learning rate is  $10^{-3}$  and is decreased by 2 at the epoch of 50, 80, 110, and 160. We use Adam [15] for optimization. We train our model on one NVIDIA QUADRO RTX 8000 GPU. In the full supervision setting, we mainly refer to existing works [5,26] to set  $w_{geo} = 100$ ,  $w_{tex} = 0.01$ ,  $w_{regu} = 0.01$ ,  $w_{jnt} = 2$ ,  $w_{vert} = 1.5$ ,  $w_{direc} = 0.06$ ,  $w_{len} = 1$ ,  $w_{pix} = 2$ ,  $w_{color} = 0.2$ ,  $w_{sil} = 0.08$ ,  $w_{ssim} = 1$ ,  $w_{perc} = 10^{-6}$ . More discussion of weight ablation studies can be seen in the supplementary materials.

### 4.2 Datasets and Evaluation Metrics

Our pipeline is initially pretrained on a synthetic dataset, either RHD or DARTset, and then train and evaluate on two real-world datasets, including FreiHAND and HO-3Dv2.

**RHD** [39]. RHD is a considerable synthetic dataset that includes 3D annotations for both singular and interacting hand postures. The dataset synthesizes hand positions from an array of 20 unique characters engaged in a diverse set of 39 actions. It offers a substantial supply of images for empirical study, with 41,258 training instances and 2,728 instances for evaluation, thereby providing a slight perturbation in the data distribution.

**DARTset** [7]. DARTset is a large-scale hand dataset with diverse poses, textures, and accessories, comprising 800K samples split into a training set (758,378) and a test set (28,877). Among these hands, 25% are assigned an accessory. In our experiment, we used 100,000 samples from the training set and 288,77 samples from the test set to pretrain our pipeline.

**FreiHAND** [40]. FreiHAND is a commonly used dataset that contains single-hand images in the wild. There are 32,560 images for training and 3,960 images

for testing. For each sample, one RGB image and annotations for 3D joints and vertices labels are provided.

**HO-3Dv2** [8]. HO-3D is 3D pose annotations for hands and objects under severe occlusions from each other. It contains annotations for 77,558 images which are split into 66,034 training images (from 55 sequences) and 11,524 evaluation images (from 13 sequences). Evaluations are conducted by submitting our estimated results to their online system.

**Evaluation Metrics.** As with existing works [5,1], we use the mean per joint position error (MPJPE) and mean per vertex position error (MPVPE) in cm between the prediction and ground truth after Procrustes alignment to evaluate the 3D reconstruction accuracy of the joints and mesh vertices. To evaluate the texture reconstruction accuracy, we use metrics that focus on the rendered image quality. The L1 distance of two images is used for a low-level representation of the reconstruction quality. SSIM [33] and PSNR are used to reflect image similarity as the rendering quality. Besides, consistent with some recent works [34,4], we also adopt LPIPS [37] as high-level metrics representing the human perception of the texture quality.

**Table 1.** Comparison on FreiHAND [40]. The column ‘‘Supv.’’ indicates whether the supervision level is 2D or 3D. The column ‘‘Tex.’’ indicates whether the method is able to reconstruct hand texture.

Methods	Supv.	MPJPE↓	MPVPE↓	Tex.	L1↓	PSNR↑	SSIM↑	LPIPS↓
Biomechanical [30]	2D	1.13	-	No	-	-	-	-
S <sup>2</sup> HAND [5]	2D	1.18	1.19	Yes	0.12	16.61	0.79	0.43
SMHR [26]	2D	1.07	1.10	Yes	-	16.64	-	-
Ours <sub>weak</sub>	2D	1.23	1.24	Yes	0.02	20.00	0.93	<b>0.09</b>
Ours <sub>self</sub>	2D	1.31	1.33	Yes	<b>0.02</b>	<b>20.04</b>	<b>0.94</b>	0.10
Ours <sub>self</sub> *	2D	1.66	1.65	Yes	0.03	17.88	0.92	0.13
Boukhayma <i>et al.</i> [1]	3D	3.50	1.32	No	-	-	-	-
ObMan [10]	3D	1.33	1.33	No	-	-	-	-
ManoCNN [40]	3D	1.10	1.09	No	-	-	-	-
ManoFit [40]	3D	1.37	1.37	No	-	-	-	-
HTML [24] <sup>†</sup>	3D	1.11	1.10	Yes	-	-	-	-
HIU [38]	3D	0.71	0.86	No	-	-	-	-
SMHR [26]	3D	0.80	0.81	Yes	-	16.64	-	-
Ours <sub>full</sub>	3D	1.21	1.23	Yes	0.03	<b>19.55</b>	0.94	0.10

<sup>†</sup>: Due to the unavailability of HTML’s 3D reconstruction implementation codes, which were not published alongside their hand layer codes, we cannot reproduce their results or provide texture reconstruction metrics for comparison.

### 4.3 Comparison to State-of-the-art Methods

In this section, we evaluate the reconstruction performance of our approach and compare it with the state-of-the-art methods on two widely used single-hand

datasets, FreiHAND [40] and HO-3Dv2 [8]. Following SMHR [26], we mainly focus on the model-based methods for a fair comparison.

**Comparison on Geometry Reconstruction Quality.** We first report the comparison results on the geometry reconstruction using MPJPE and MPVPE with state-of-the-art methods [30,5,26,1,10,40,24,38]. Note that most existing works focus on the 3D geometry reconstruction and ignore textures, while our work is able to reconstruct the geometry as well as the high-quality texture.

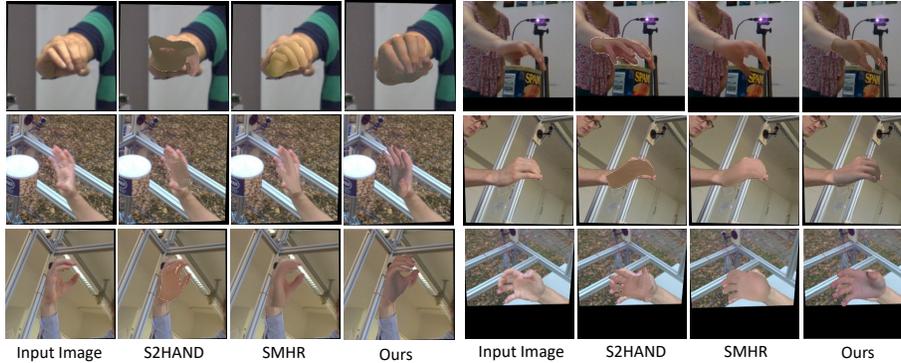
Tab. 1 shows the results on FreiHAND [40], and Tab. 2 shows the results on HO-3Dv2 [8]. Ours<sub>full</sub> is our method under full 3D supervision, *i.e.*, supervised with ground truth 3D joints and vertices. Ours<sub>weak</sub> is weakly supervised by only 2D ground truth joints, *i.e.*,  $\mathcal{L}_{jnt}^{2D}$ . While the definition of self-supervision is controversial, we implement two versions for discussion. Ours<sub>self</sub> follows S<sup>2</sup>HAND [5] to detect noisy 2D joints via OpenPose [28] and utilize the confidence-aware 2D joint loss, and another version, Ours<sub>self\*</sub>, does not adopt any form of 2D annotation and is only supervised by input images using texture reconstruction consistency losses.

**Table 2.** Comparison on HO-3Dv2 [8]. The column ‘‘Supv.’’ indicates whether the supervision level is 2D or 3D. The column ‘‘Tex.’’ indicates whether the method is able to reconstruct hand texture.

Methods	Supv.	MPJPE↓	MPVPE↓	Tex.	L1↓	PSNR↑	SSIM↑	LPIPS↓
PeCLR [29]	2.5D	1.09	-	No	-	-	-	-
S <sup>2</sup> HAND [5]	2D	1.14	1.12	Yes	-	13.92	-	-
SMHR [26]	2D	1.03	1.01	Yes	-	16.78	-	-
Ours <sub>weak</sub>	2D	1.16	1.23	Yes	0.01	23.66	0.95	0.12
Ours <sub>self</sub>	2D	1.23	1.26	Yes	<b>0.01</b>	<b>26.50</b>	<b>0.96</b>	<b>0.10</b>
Ours <sub>self*</sub>	2D	1.48	1.51	Yes	-	-	-	-
HO3D [8]	3D	1.07	1.06	No	-	-	-	-
ObMan [10]	3D	-	1.10	No	-	-	-	-
Photometric [9]	3D	1.11	1.14	No	-	-	-	-
SMHR [26]	3D	1.01	0.97	Yes	-	16.78	-	-
Ours <sub>full</sub>	3D	1.16	1.22	Yes	0.01	<b>23.97</b>	0.95	0.12

Our method achieves comparable pose and shape accuracy with state-of-the-art methods under all kinds of supervision. Note that previous works like S<sup>2</sup>HAND [5] and SMHR [26] use additional branches to perform keypoint estimation or hand localization while our pipeline does not rely on an extra branch. Our focus lies in studying the impact of texture reconstruction on pose and shape learning across different levels of supervision, rather than solely aiming for the best results in pose and shape reconstruction. Consequently, we employ a simplified structure and training settings. Therefore, our experimental results remain competitive, demonstrating the robustness and effectiveness of our approach.

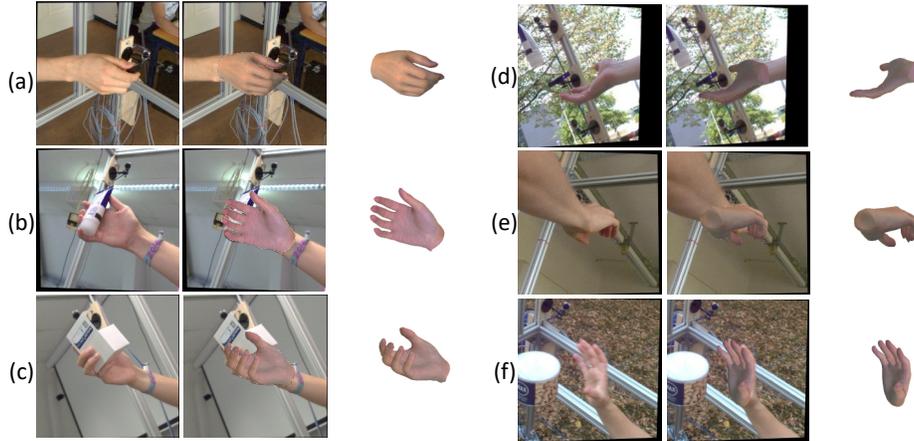
**Comparison on Texture Reconstruction Quality.** Secondly, we compare our texture reconstruction quality with several existing methods that are able to generate hand textures [5,26]. (See the texture metrics in Tab. 1 and Tab. 2.)



**Fig. 3.** Visualization comparison of texture quality with the state-of-the-art methods on FreiHAND and HO-3Dv2. The results clearly demonstrate the superior accuracy, high fidelity, and overall plausibility of our reconstructed hand texture compared to other methods.

It can be clearly seen that our texture reconstruction quality outperforms all the methods. It achieves a 20.2% increase on FreiHAND and a 57.9% increase on HO-3Dv2 in PSNR compared to the previous state-of-the-art method, SMHR [26]. This is mainly because of our high-fidelity texture representation, *i.e.*, physical-based rendering [23], which represents texture using multiple high-resolution UV maps. Instead, S<sup>2</sup>HAND [5] and SMHR [26] simply represent textures as per-vertex RGB values, which are crude and lack realism. Moreover, we achieve high LPIPS values, showing that our method produces high-fidelity hands in human perception. Here we exclude the texture quality results of Self\* from the HO-3Dv2 evaluation because of the significant influence of inaccurate pose and shape estimation on these results.

Fig. 4 showcases the visualization results of our reconstructed hands, highlighting the generation of high-fidelity hands with detailed skin features such as palm prints, nails, and bone bumps. Even in challenging scenarios with extreme poses and severe occlusions, our method produces plausible hand reconstructions despite limited information about the hand texture. Besides, Fig. 3 visually compares our method with state-of-the-art approaches, demonstrating the superior accuracy, high fidelity, and plausibility of our reconstructed hand texture.



**Fig. 4.** Qualitative hand reconstruction results on FreiHAND and HO-3Dv2. (a)(b) High-fidelity details such as skin bumps and palm prints. (c)(d)(e) Plausible hands even severely occluded, in extreme poses or viewpoints. (f) Plausible hand even when the input image is in motion blur.

#### 4.4 Evaluation on Texture Reconstruction Consistency with Different Supervision Settings

In this section, we evaluate the effectiveness of our proposed texture reconstruction consistency under different levels of supervision. The experiments in Tab. 3 align with the settings of  $Ours_{full}$ ,  $Ours_{weak}$ ,  $Ours_{self}$ , and  $Ours_{self^*}$ , respectively. In experiments without texture reconstruction, we do not implement texture reconstruction consistency losses. Since  $Ours_{self^*}$  solely employs texture reconstruction losses, there is no corresponding experiment without texture.

Under full 3D supervision, the reconstruction of textures does not significantly affect the pose and shape results. This is attributed to the stronger influence of full 3D supervision compared to textures, causing the texture reconstruction less effective. However, when we reduce the level of supervision to weak 2D supervision, reconstructing textures does not attribute to the pose and shape accuracy. In fact, it introduces noise, resulting in a marginal increase of 0.05 in MPJPE and 0.07 in MPVPE. Nevertheless, as we further decrease the level of supervision to self-supervision using detected noisy keypoints, the texture reconstruction consistency becomes instrumental in learning accurate pose and shape, which improves results in a reduction of MPJPE by 0.02 and MPVPE by 0.03. This indicates that texture consistency loss improves hand reconstruction in the absence of reliable annotations, reducing the reliance on labeled data. In the last experiment,  $Self^*$ , we adopt a more rigorous form of self-supervision with solely the self-consistency of textures, without any explicitly defined keypoints as auxiliaries for the training. However, this approach yields unsatisfactory results, with an MPJPE of only 1.66. We believe this is be-

cause hand has complex pose configurations and depth ambiguities. Only using texture consistency loss can not learn this information well, which focuses more on pixel-level and perceptual-level features. This also indicates that using only the proposed texture consistency loss terms is insufficient to extract adequate information about hand pose and shape from the images.

**Table 3.** Effect of texture reconstruction consistency with different supervision settings on FreiHAND [40]. The first column indicates the supervision setting, and the second column indicates whether texture reconstruction is included.

Supervision	Texture	MPJPE↓	MPVPE↓	L1↓	PSNR↑	SSIM↑	LPIPS↓
Full	Yes	1.21	1.23	0.03	19.40	0.94	0.11
Full	No	1.21	1.24	-	-	-	-
Weak	Yes	1.28	1.31	0.02	20.00	0.94	0.10
Weak	No	1.23	1.24	-	-	-	-
Self	Yes	1.31	1.33	0.02	20.04	0.94	0.10
Self	No	1.33	1.35	-	-	-	-
Self*	Yes	1.66	1.65	0.03	17.88	0.92	0.13

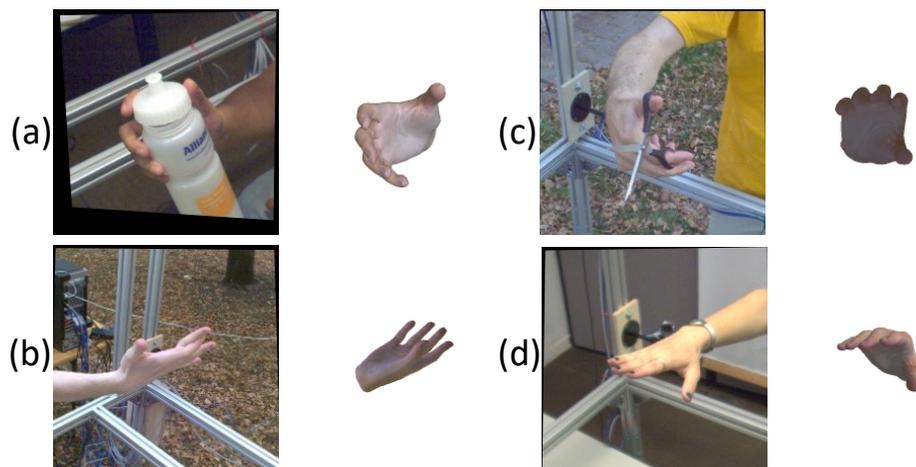
#### 4.5 Ablation Study

**Effect of Our Components.** We evaluate the effect of our components in Tab. 4, using the baseline model with full 3D supervision. The proposed light estimator plays a crucial role in estimating directional lighting and disentangling it from texture, contributing to an increase in PSNR. Besides, replacing the EfficientNet-b3 backbone with ResNet50 [11] leads to a significant decrease in pose and shape performance. Furthermore, we explore pretraining the pose and shape estimator on DART [7] instead of RHD [39], which leads to a slight increase in performance. DART offers a significantly larger dataset, with more diverse pose and shape distribution than RHD. Here we only use 1/7 of its training dataset for pretraining, so the result highlights its potential for enhancing accuracy. Note that we chose RHD as the pretraining dataset for our baseline to ensure a fair comparison, as it is the most widely utilized pretraining dataset. We also assess the effectiveness of our losses. Removing  $\mathcal{L}_{vert}$  in full supervision results in a slight decrease in MPVPE. Removing  $\mathcal{L}_{perc}$  leads to a noticeable decrease in the quality of the reconstructed texture. Additional ablation studies on losses can be found in the supplementary materials.

**Limitations.** Some limitations and weaknesses remain. First, using directional light to represent illumination is limited in some complex scenarios. Furthermore, our method does not consider hand-object interaction scenarios. Fig. 5 shows some failure cases, including object occlusion, extreme pose and texture, and personal features. To address these limitations, we suggest exploring an enhanced pipeline and considering hand-object interaction scenarios.

**Table 4.** Ablation study on FreiHAND test set [40]. We study the impact by changing one specific component in each experiment.

Methods	MPJPE↓	MPVPE↓	PSNR↑
Ours	1.21	1.23	19.547
w/o light estimator	1.22	1.25	18.983
ResNet50 as backbone	1.30	1.32	19.582
pretrain on DART	1.20	1.22	19.755
w/o $\mathcal{L}_{vert}$	1.21	1.25	-
w/o $\mathcal{L}_{perc}$	1.21	1.23	19.512

**Fig. 5.** Failure cases. (a) Incorrect texture color due to occlusion. (b) Incorrect pose due to extreme pose. (c) Incorrect pose due to severe hand-object interaction. (d) Unable to reconstruct the black nail.

## 5 Conclusions

We present a 3D hand reconstruction method that is able to generate hands with high-fidelity textures from single input images. Our approach surpasses current state-of-the-art methods in texture quality while maintaining competitive pose and shape accuracy. Through extensive experiments across various levels of supervision, we provide valuable insights into the influence of high-fidelity hand texture reconstruction on pose and shape estimation: in scenarios with noisy and relatively weak supervision, textures effectively contribute to pose and shape learning, whereas their impact becomes less significant under stronger supervision. In future work, we aim to extend our research to more complex scenarios that involve hand-object interaction and explore advanced loss terms for texture reconstruction consistency.

## References

1. Boukhayma, A., Bem, R.d., Torr, P.H.: 3d hand shape and pose from images in the wild. In: *Computer Vision and Pattern Recognition*. pp. 10843–10852 (2019)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: *Computer Vision and Pattern Recognition*. pp. 7291–7299 (2017)
3. Chen, W., Gao, J., Ling, H., Smith, E., Lehtinen, J., Jacobson, A., Fidler, S.: Learning to predict 3d objects with an interpolation-based differentiable renderer. In: *Advances in Neural Information Processing Systems* (2019)
4. Chen, X., Wang, B., Shum, H.Y.: Hand avatar: Free-pose hand animation and rendering from monocular video. *arXiv preprint arXiv:2211.12782* (2022)
5. Chen, Y., Tu, Z., Kang, D., Bao, L., Zhang, Y., Zhe, X., Chen, R., Yuan, J.: Model-based 3d hand reconstruction via self-supervised learning. In: *Computer Vision and Pattern Recognition* (2021)
6. Corona, E., Hodan, T., Vo, M., Moreno-Noguer, F., Sweeney, C., Newcombe, R., Ma, L.: Lisa: Learning implicit shape and appearance of hands. In: *Computer Vision and Pattern Recognition*. pp. 20533–20543 (2022)
7. Gao, D., Xiu, Y., Li, K., Yang, L., Wang, F., Zhang, P., Zhang, B., Lu, C., Tan, P.: Dart: Articulated hand model with diverse accessories and rich textures. *arXiv preprint arXiv:2210.07650* (2022)
8. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: *Computer Vision and Pattern Recognition* (2020)
9. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: *Computer Vision and Pattern Recognition*. pp. 571–580 (2020)
10. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: *Computer Vision and Pattern Recognition*. pp. 11807–11816 (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* (2015)
12. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headerf: A real-time nerf-based parametric head model. In: *Computer Vision and Pattern Recognition* (2022)
13. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: Neuman: Neural human radiance field from a single video. In: *European Conference on Computer Vision*. pp. 402–418. Springer (2022)
14. Karunratanakul, K., Prokudin, S., Hilliges, O., Tang, S.: Harp: Personalized hand reconstruction from a monocular rgb video. *arXiv preprint arXiv:2212.09530* (2022)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **60**(6), 84–90 (2017)
17. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: *Computer Vision and Pattern Recognition*. pp. 4681–4690 (2017)
18. Li, Y., Zhang, L., Qiu, Z., Jiang, Y., Li, N., Ma, Y., Zhang, Y., Xu, L., Yu, J.: Nimble: A non-rigid hand model with bones and muscles. *ACM Transactions on Graphics* **41**(4) (jul 2022)

19. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* **34**(6), 248:1–248:16 (Oct 2015)
20. Moon, G., Shiratori, T., Lee, K.M.: Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In: *European Conference on Computer Vision* (2020)
21. Mundra, A., Wang, J., Habermann, M., Theobalt, C., Elgharib, M., et al.: Livehand: Real-time and photorealistic neural hand rendering. *arXiv preprint arXiv:2302.07672* (2023)
22. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: *NeurIPS Autodiff Workshop* (2017)
23. Pharr, M., Jakob, W., Humphreys, G.: 01 - introduction. In: Pharr, M., Jakob, W., Humphreys, G. (eds.) *Physically Based Rendering (Third Edition)*, pp. 1–55. Morgan Kaufmann, Boston, third edition edn. (2017)
24. Qian, N., Wang, J., Mueller, F., Bernard, F., Golyanik, V., Theobalt, C.: HTML: A Parametric Hand Texture Model for 3D Hand Reconstruction and Personalization. In: *European Conference on Computer Vision*. Springer (2020)
25. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501* (2020)
26. Ren, J., Zhu, J., Zhang, J.: End-to-end weakly-supervised single-stage multiple 3d hand mesh reconstruction from a single rgb image. *Computer Vision and Image Understanding* **232**, 103706 (2023)
27. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.* **36**(6) (nov 2017)
28. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: *Computer Vision and Pattern Recognition* (2017)
29. Spurr, A., Dahiya, A., Wang, X., Zhang, X., Hilliges, O.: Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In: *International Conference on Computer Vision*. pp. 11230–11239 (2021)
30. Spurr, A., Iqbal, U., Molchanov, P., Hilliges, O., Kautz, J.: Weakly supervised 3d hand pose estimation via biomechanical constraints. In: *European Conference on Computer Vision*. pp. 211–228. Springer (2020)
31. Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In: *Advances in Neural Information Processing Systems* (2021)
32. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*. pp. 6105–6114. PMLR (2019)
33. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004)
34. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In: *Computer Vision and Pattern Recognition*. pp. 16210–16220 (June 2022)
35. Xu, H., Alldieck, T., Sminchisescu, C.: H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In: *Advances in Neural Information Processing Systems*. vol. 34, pp. 14955–14966 (2021)

36. Xu, T., Fujita, Y., Matsumoto, E.: Surface-aligned neural radiance fields for controllable 3d human synthesis. In: *Computer Vision and Pattern Recognition*. pp. 15883–15892 (2022)
37. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Computer Vision and Pattern Recognition*. pp. 586–595 (2018)
38. Zhang, X., Huang, H., Tan, J., Xu, H., Yang, C., Peng, G., Wang, L., Liu, J.: Hand image understanding via deep multi-task learning. In: *International Conference on Computer Vision*. pp. 11281–11292 (2021)
39. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. Tech. rep., arXiv:1705.01389 (2017)
40. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: *International Conference on Computer Vision*. pp. 813–822 (2019)