Handy: Towards a high fidelity 3D hand shape and appearance model

Rolandos Alexandros Potamias[1] Stylianos Ploumpis[1] Stylianos Moschoglou[1]
Vasileios Triantafyllou[2] Stefanos Zafeiriou[1]
[1]Imperial College London [2]Cosmos Designs Ltd
{r.potamias,s.ploumpis,s.moschoglou,s.zafeiriou}@imperial.ac.uk

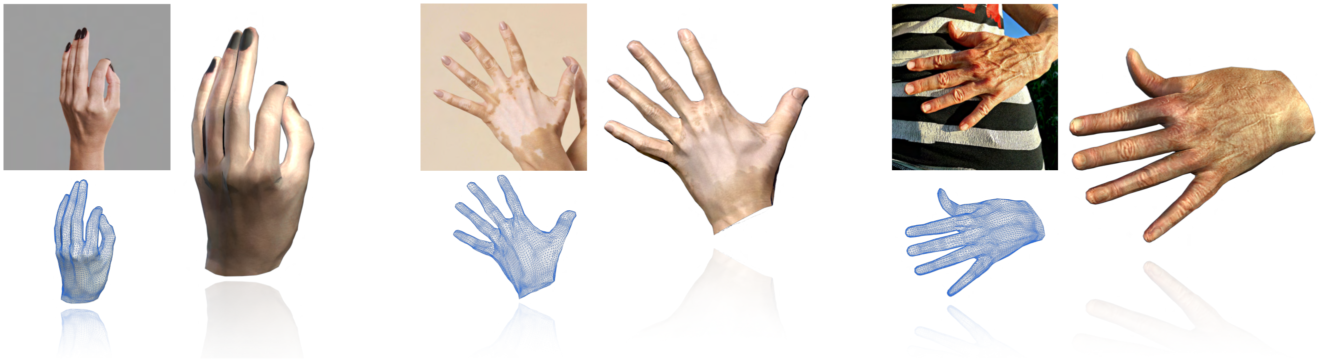

Figure 1. Our proposed hand model is able to generalise and accurately reconstruct the 3D hand shape and appearance from a single in-the-wild image. High frequency details are visible in our reconstructions such as wrinkles, veins, nail polish etc.

Abstract

Over the last few years, with the advent of virtual and augmented reality, an enormous amount of research has been focused on modeling, tracking and reconstructing human hands. Given their power to express human behavior, hands have been a very important, but challenging component of the human body. Currently, most of the state-of-the-art reconstruction and pose estimation methods rely on the low polygon MANO model. Apart from its low polygon count, MANO model was trained with only 31 adult subjects, which not only limits its expressive power but also imposes unnecessary shape reconstruction constraints on pose estimation methods. Moreover, hand appearance remains almost unexplored and neglected from the majority of hand reconstruction methods. In this work, we propose "Handy", a large-scale model of the human hand, modeling both shape and appearance composed of over 1200 subjects which we make publicly available for the benefit of the research community. In contrast to current models, our proposed hand model was trained on a dataset with large diversity in age, gender, and ethnicity, which tackles the limitations of MANO and accurately reconstructs out-of-distribution samples. In order to create a high quality texture model, we trained a powerful GAN, which preserves high frequency details and is able to generate high resolution hand textures. To showcase the capabilities of the proposed model, we built a synthetic dataset of textured hands and trained a hand pose estimation network to reconstruct both the shape and appearance from single images. As it is demonstrated in an extensive series of quantitative as well as qualitative experiments, our model proves to be robust against the state-of-the-art and realistically captures the 3D hand shape and pose along with a high frequency detailed texture even in adverse "in-the-wild" conditions.

1. Introduction

Humans express their emotions mainly using their facial expressions and hands. Hand movements and poses are strong indicators of body language and can convey meaningful messages which can be key factors in human behavioral analysis. For this, hands have been widely studied in regard to their biometric applications [11, 40]. 3D hand models lead the technological developments of crucial tasks for virtual reality such as human hand tracking

Project page: https://github.com/rolpotamias/handy.

[19, 36, 42, 51] and pose estimation [18, 55]. Specifically, hand pose estimation algorithms utilize these models in order to reconstruct a subject's hand from a monocular depth or RGB image. However, most of the current state-of-the-art methods on 3D hand reconstruction and pose estimation rely on low polygon models, with minimum diversity in terms of age, gender, and ethnicity and without any hand texture appearance [39].

In particular, MANO [39] is considered the most popular hand model, which pioneered the construction of a parametric human hand model. Apart from its low polygon resolution (778 vertices), it is only composed of 31 subjects, which limits the accuracy of high fidelity 3D reconstruction methods. A statistical model with such a low number of samples will always constrain the reconstruction of hand shapes of diverse age and ethnicity groups. In the same context, despite the efforts of implementing strong pose priors to accurately constrain parametric models on valid hand poses [31], reconstruction methods are still dependent on a limited shape model. Importantly, current parametric models are constructed only by adults' hand shapes in the age range of 20-60 years old, disregarding the shape variations out of this age range. We experimentally show that children's hands significantly differ in terms of shape from adults' hands, which makes current shape models prone to reconstruction errors.

Additional to the shape component, a major limitation of current hand models is the absence of a high resolution texture model. Despite the necessity in virtual and augmented reality for a personalized appearance reconstruction, there are only a few studies that attempted to model hand texture along with shape and pose. In particular, current methods on hand texture reconstruction from monocular images are constrained on limited demographic variations and low resolution textures that are ill-suited for real-world applications [8–10, 37, 47]. Recently, HTML [37] proposed the largest available parametric texture model of the human hand composed of 51 subjects. Given that the texture component is based on Principal Component Analysis (PCA) of low resolution texture UV maps, the generated textures tend to be blurry, lacking the high frequency details of the hand. Low resolution textures not only limit the fidelity of RGB reconstructions but also the generations of realistic synthetic data. Currently, state-of-the-art hand-object detection methods [3, 16, 49] train their models on synthetic datasets with low resolution textures such as HTML or vertex colors, which subsequently constrain the quality of the resulting reconstructions.

In this study, we propose "Handy", the first large-scale parametric shape and texture hand model composed of over 1200 subjects. Given these high resolution textured scans with large demographic, gender, and age variations, we built a high resolution hand model that overcomes the shape limitations of previous state-of-the-art models. To the best of our knowledge, this is the first hand model that captures subjects with ages from 1 to 81 years old. The scans come with high resolution textures which enable the creation of a highly detailed texture model. In contrast to HTML [37], we built a high resolution texture model by using a style-based GAN which allows modeling high frequency details of the human hand (e.g., wrinkles, veins, nail polish). Under a series of experiments we show that the proposed parametric model overcomes the limitations of previous methods and we present the first, to the best of our knowledge, high fidelity texture reconstruction method from single "in-the-wild" images.

In particular, besides the success of 3D hand reconstruction from monocular depth and RGB images, there are currently only a few methods that are able to reconstruct the pose along with the shape and texture components. Existing 3D hand datasets only contain hand annotations in terms of pose and global rotation and they usually neglect hand shape variations by modeling only a mean hand shape. Additionally, the lack of ground-truth high resolution texture maps limits current hand reconstruction methods to properly predict the appearance of a given hand. To enable texture modeling, we follow the trend of synthetic data generation, and we have built a large-scale dataset containing annotations in terms of pose, shape, and texture information. To summarize, the contributions of our work are the following:

- We make publicly available a large-scale shape and appearance model of the human hand for the benefit of the research community, built by over 1200 3D hands scans with a large diversity in age, gender, and ethnicity.
- We create a synthetic dataset for monocular 3D hand reconstruction given our high fidelity hand model and make it publicly available. As shown in the experimental section, our synthetic dataset aid off-the-shelf reconstruction methods to improve results.
- We present a high fidelity appearance reconstruction method from monocular images which is able to reconstruct high frequency details such as wrinkles, veins, nail polish, etc.

2. Related work

Parametric Hand Models. Over the years, several hand models have been proposed in the literature to approximate hand articulations. Initially, Oikonomidis *et al.* [33] attempted to model hand shape as a collection of geometric primitives such as elliptic cylinders, ellipsoids, spheres, and cones. Following [33], various approaches were proposed to model hand joints using anisotropic Gaussians [44], a collection of spherical meshes [46], or a union of convex bodies [28]. Schmidt *et al.* [41] proposed the first implicit representation of the articulated hand using the popular Signed Distance Function. Khamis *et al.* [23] proposed the first linear blend skinning (LBS) model constructed from 50 subject scans. The authors modeled hand poses and shape variations using a low dimensional PCA. To tackle the volume loss and restrict unrealistic poses of the LBS, Romero *et al.* [39] learned pose dependent corrective blend shapes from the scans of 31 subjects and proposed the MANO parametric model. Li *et al.* [26] proposed the NIMBLE to model the interior of the hand, i.e. bones and muscles. Recently, HTML [37] attempted to create a parametric appearance model of the human hand by collecting hand textures from 50 subjects. However, given the limited amount of data, the authors train a PCA model on the UV space resulting in low resolution textures. To address the aforementioned limitations, the present work proposes the first large-scale model of both hand shape and appearance of the human hand, composed of over 1200 scans.

Hand pose estimation. 3D hand pose estimation has been a long studied field, originally tackled by deforming a hand model to volumetric [17] and depth images [15,32,43,45]. Initially, 3D pose estimation was considered as a fitting problem where a 3D parametric model was used to fit 2D keypoints [34,35]. De La Groce *et al.* [9] pioneered hand pose tracking from single RGB images by solving an optimization problem. The advent of deep learning methods has shifted the research interest to sparse joint keypoints prediction from RGB images using Convolutional Neural Networks (CNNs) [18, 30, 34, 54]. Most of such methods attempt to directly predict dense 3D hand positions by regressing the MANO model [39] parameters [2,3,53], which constrain them to the shape and pose space of MANO. Several methods deviated from MANO's parameter space, by directly regressing 3D vertex positions using graph neural networks [13, 24, 25]. Hasson *et al.* [16] proposed a CNN-based method that regresses MANO and AtlasNet parameters to reconstruct 3D hand poses together with various object shapes. Recently, a handful of methods attempted to reconstruct objects along with hands by using implicit [22, 50], parametric [4, 49] or a combination of both representations [7].

Synthetic datasets for hand pose estimation. Synthetic datasets have been proven very effective, boosting training performance and overcoming data limitations in many applications, ranging from face reconstruction [48] to pedestrian detection [12]. Numerous amount of hand pose methods have been trained using synthetic data generated under different hand poses and illumination environments [3,5,13,49,54]. Hasson *et al.* [16] rendered synthetic data using the SMLP model [27] under various hand poses from the GraspIt dataset [29]. Apart from the hands, the authors used objects from ShapeNet to generate a dataset of hand-object interactions. However, all of the aforementioned studies are limited to only a few texture variations [54] or low resolution hand textures [3, 16, 47, 49], creating a domain gap between synthetic and real-world images. To alleviate such domain gap, we propose a new dataset of both hands and objects, similar to [16], using high resolution textures of hands, making a step towards a photorealistic synthetic hand dataset.

3. Handy: Shape and Appearance Model

In this Section, we introduce our large-scale hand model "Handy". We begin by introducing the 3D dataset we collected with which we built our high fidelity shape and texture model. Then, we describe how we brought into dense correspondence the entire hand dataset and created our large-scale shape model. Finally, we explain how we trained a style-based appearance model which preserves all the high frequency details of the human hand.

3.1. Large-scale 3D hand dataset

As mentioned in Section 1, we collected a large dataset comprising of textured 3D hand scans. Our hand data were captured during a special exhibition at the Science Museum, London. The capturing apparatus utilized for this task was a 3dMD 4 camera structured light stereo system, which produces high quality dense meshes. The raw scans have a resolution of approximately 30,000 vertices. We captured a total of 1208 distinct subjects with available metadata about them, including their gender (53% male, 47% female), age ($1-81$ years old), height ($80-210$ cm), and ethnicity (82% White, 9% Asian, 7% Mixed and 2% black), as shown in Figure 2.

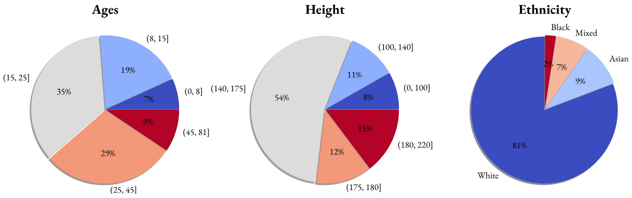

Figure 2. Distribution of demographic characteristics of the scanned subjects. The collected hand dataset covers a large variety of ages, heights, and ethnicities.

Most notably, our collected hand scans exhibit a large diversity in terms of age, ethnicity, and height, which provide a step towards a universal hand model. Compared to previous methods [37, 39], the scans collected include over 360 children aged less than 12 years old and 100 elderly subjects aged over 60 years old. In order to capture different pose variations, each subject was instructed to perform a range of hand movements according to a specific protocol each day for a period of 101 days. Some example images can be seen in Figure 7. Since our dataset contains personal data from human subjects, we have ensured that the collection of such data has been carefully conducted in accordance with the ethics guidelines.

3.2. Shape model construction

In order to create a statistical shape model of the human hand, we begin by rigidly aligning a set of 3D scans with a common template mesh. We utilize two different resolution templates in terms of polygon count for our method. As a low polygon resolution template, we utilize the MANO template with 778 vertices, which can be directly adapted to SMPL model [27]. For high quality hand modeling, we utilize a high resolution hand template in terms of polygons, comprising of 8407 vertices. In order to bring the raw scans into dense correspondence, we render them from multiple views and use MediaPipe framework [51] to detect 2D joint locations. We lift the 2D joint locations to 3D by utilizing a linear triangulation and then detect the fingertips by using the projection of the finger skeleton to the tips of the surface. Subsequently, we use the lifted 3D landmarks to fit/align the pose of our template to the 3D scan surface. Finally, in order to acquire the final hand dense registrations, we apply the Non-rigid Iterative Closest Point algorithm (NICP) [1] between our hand template mesh and the 3D raw scans.

After having all the 3D raw hand scans into dense correspondence with our high resolution template, we normalize them to a canonical open-palm pose in order to avoid capturing any unnecessary deformations into our final shape model. We construct a deformable hand shape model described as a linear basis of shapes. In particular, using PCA, we build a hand model with N vertices that is described by an orthonormal basis after keeping the first n_c principal components $\mathbf{U} \in \mathbb{R}^{3N \times n_c}$ and their associated λ eigenvalues. New hand instances can then be generated by regressing the shape parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_{n_c}] \in \mathbb{R}^{n_c}$ as:

$$B_s(\boldsymbol{\beta}) = \mathbf{T} + \sum_{i=0}^{n_c} \mathbf{U_i}\beta_\mathbf{i} \in \mathbb{R}^{3N} \tag{1}$$

where $\mathbf{T} \in \mathbb{R}^{3N}$ refers to the mean hand shape. Variations of the first 5 shape components are illustrated in Figure 3.

Finally, the articulated hand model can be defined as:

$$\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}) = W(T_p(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \boldsymbol{\mathcal{W}}) \tag{2}$$

$$T_p(\boldsymbol{\beta}, \boldsymbol{\theta}) = B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) \tag{3}$$

where $W(\cdot)$ corresponds to a linear blend skinning function (LBS) that is applied to the articulated hand mesh with posed shape T_p, $\boldsymbol{\mathcal{W}}$, J define the blend weights and kinematic tree of joint locations, respectively, and $\boldsymbol{\beta}, \boldsymbol{\theta}$, are the shape and pose parameters, respectively. To tackle the joint collapse of typical LBS function, we follow MANO [39] and use the learned pose corrective blendshapes that produce more realistic posed hands:

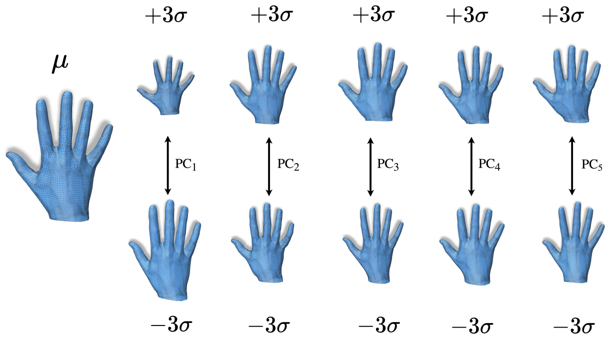

Figure 3. Mean shape $\mathbf{T}$ and the first five principal components, each visualized as additions and subtractions of 3 standard deviations ($\pm 3\sigma$) away from the mean shape.

$$B_p(\boldsymbol{\theta}) = \sum_{i=0}^{9K} (R_i(\boldsymbol{\theta}) - R_i(\boldsymbol{\theta}^*))\, \mathbf{P}_i \tag{4}$$

where $\mathbf{P}_i$ are the pose blend shapes, K is the number of joints of the hand model, $R_i(\boldsymbol{\theta})$ is a function that maps pose parameters $\boldsymbol{\theta}$ to the rotation matrix of joint i and $\boldsymbol{\theta}^*$ refers to the canonical pose.

3.3. High resolution appearance model

As also shown in HTML [37], in order to train a texture model, the hand scans need to be brought into correspondence. To achieve this optimally, we asked a graphics artist to design a UV hand template and used it as a reference template to unwrap the scans. However, the hand scans were acquired using constrained light conditions with baked shadows. As a result, before carrying out any training procedure, we followed a pre-processing step on the UV textures to remove the shading and illumination. In particular, we applied PCA to the UV textures and identified the components that mostly describe the shading factors. We then subtracted those components from each texture UV map to remove their unnecessary shading. Finally, we followed an image processing step that mapped hand textures to more natural colors, which entailed increasing the brightness, gamma correction, and slightly adjusting the hue value.

For the training process, rather than modeling the appearance space in a low frequency PCA domain as other methods do [37], we utilized a powerful GAN architecture [20] to model the hand textures. Given the limited number of collected data, we used a smaller learning rate of 0.001. We also found that a regularization weight γ of 50 further assisted in the FID score as well as the visual quality of the final results. In Figure 4, we showcase some random generations of our proposed high fidelity appearance model. Utilizing such GAN architecture, we are capable of preserving the high frequency skin details whilst avoiding any smoothness created by a PCA model. Qualitative results of our texture reconstruction in Figure 8 can validate this premise.

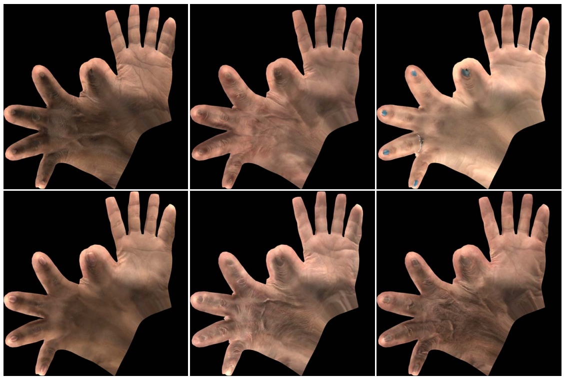

Figure 4. Generated high quality texture UV maps from our GAN appearance model.

4. Experiments

4.1. Intrinsic Evaluation of Proposed Hand Model

In this Section, we evaluate the proposed hand shape model and compare it with the MANO model [39] that is commonly used in the literature. We follow the common practice and compare the two shape models in terms of *compactness, generalization* and *specificity*. The two models were tested on the test split of MANO dataset. We also report a variation of the proposed model that utilizes the MANO template as described in Section 3.2. Note that only the first 10 out of 31 principal components of MANO are publicly available.

Compactness. In Figure 5 (left) we report the compactness of the two models. Compactness describes the percentage of the variance of the training dataset explained by the model given a number of the retained principal components. Figure 5 (left) illustrates that the proposed model better explains the dataset variations, breaking the threshold of 90% variance from the 5th component compared to the MANO model that reaches 90% variance at the 9th component.

Generalization demonstrates the ability of the model to generate new hand instances that were not present in the training set. We measure the generalization error as the mean per-vertex distance of each mesh on the MANO test set and its corresponding model re-projection. Figure 5 (middle) reveals that the proposed model achieves better out-of-distribution generalization and lower standard deviation compared to MANO model which fails to generate novel hand shapes.

	Age <8	Age <12
MANO [39]	0.78	0.77
Proposed w/ MANO ($n_c = 10$)	0.48	0.44
Proposed w/ MANO ($n_c = 30$)	0.28	0.25
Proposed ($n_c = 10$)	0.44	0.42
Proposed ($n_c = 30$)	**0.24**	**0.21**

Table 1. Per vertex reconstruction error on 20 children's hands in mm. We also report the performance of the proposed model with the MANO template (w/MANO) and use a different number of components (n_c) for a fair comparison. Bold denotes the best performance.

Specificity. Finally, we report the specificity error, which measures the realism of the generated hand shapes and their similarity to the training samples. In particular, we generated 1,000 hand shapes and measured their per-vertex distance from the closest sample on the ground-truth datasets. For a fair comparison, the samples used to train each model serve as ground-truth shapes. Figure 5 (right) shows that the proposed method exhibits better specificity error compared to the MANO model by approximately 2.5mm. Note that the slight deviation between the Proposed and the Proposed w/MANO models is caused by the high resolution (8704 vertices) of the proposed hand template which leads to more detailed shapes compared to the MANO template (778 vertices).

4.2. Reconstruction of children's hands

A major limitation of current state-of-the-art hand models is that they were trained using limited data from specific age groups that do not reflect real hand variations. In particular, we examined the reconstruction error of 20 children's hands below the age of 12 that were not present in the training set. Table 1 highlights the reconstruction capabilities of the proposed hand model that was built with 1208 subjects with diverse age groups compared to the commonly used MANO model, which is composed of only 31 adult hands. Figure 6 shows the color-coded per vertex error which validates the superiority of the proposed model in children's hands reconstructions. As expected, MANO model fails to properly reconstruct the main anatomical difference between adults' and children's hands, which mostly lies on the back of the hand.

4.3. 3D Reconstruction from single images

Following the pathway of many hand pose estimation methods, we create a synthetic dataset to train our hand reconstruction model. In particular, we generated 30,000 texture images from the GAN model to curate a synthetic dataset with textured hands. To increase the realism of the synthetic data, similar to [16], we render hands that interact with objects of the ShapeNet dataset and we complete the

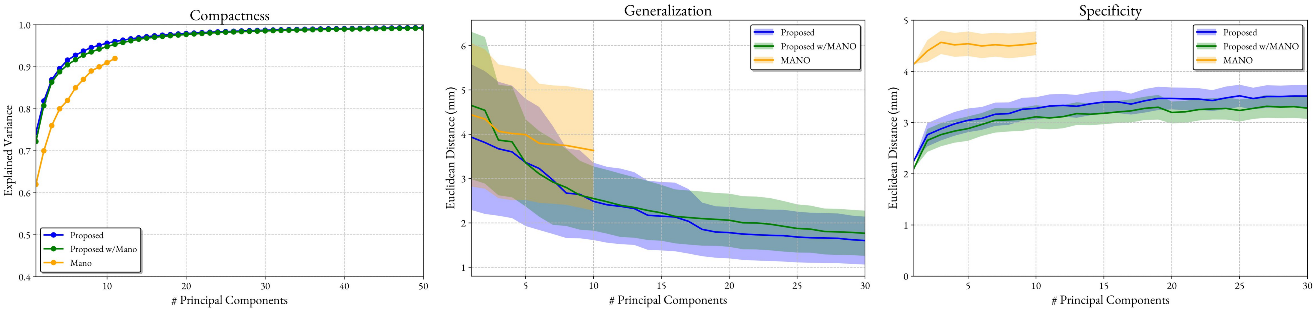

Figure 5. Evaluation of compactness, generalization and specificity against MANO model.

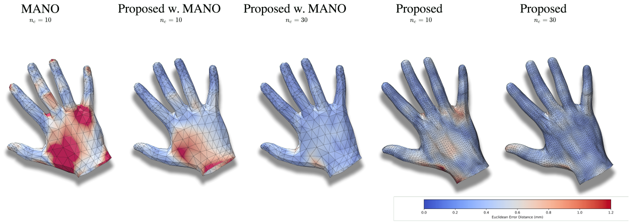

Figure 6. Color coded reconstruction error of children's hands.

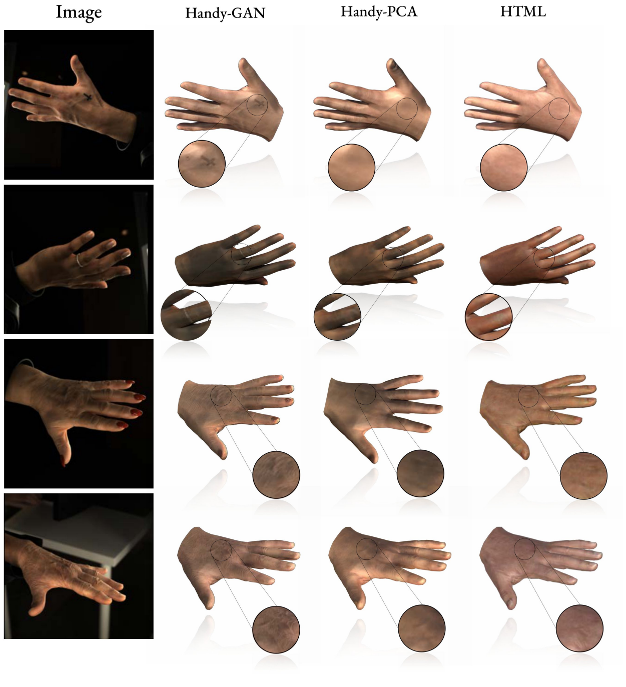

Figure 7. Hand shape and appearance reconstructions from single images under controlled conditions.

hand shape with bodies from SMPL model. In contrast to the Obman dataset [16], we use high resolution hand textures that bridge the domain gap between synthetic and "in-the-wild" hand images. We use several illuminations, lighting, and camera configurations to create diverse synthetic renderings. In order to leverage the proposed Handy model, we modified an off-the-self method [3, 14, 16], by substituting the MANO parametric model with Handy. Such method includes a ResNet50 image encoder, pre-trained on ImageNet and two branches that regress the pose and shape latent codes of our parametric model. Unlike previous methods, in order to also reconstruct the texture, we add two extra branches that regress the latent space $\mathbf{w}$ of the texture model and the camera configuration $(s, \mathbf{t})$. We train our network using a set of loss functions that enable accurate hand pose, shape, and appearance estimation. In particular, similar to [3, 16, 55], we enforce shape and pose estimation by applying a loss on both the latent shape and pose parameters and the generated 3D vertex positions:

$$\mathcal{L}_{\beta} = \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2, \quad \mathcal{L}_{\theta} = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2 \\ \mathcal{L}_{3D} = \sum_i \|\mathbf{v}_i - \hat{\mathbf{v}}_i\|_2 \tag{5}$$

where $\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{v}$ denote the predicted shape, pose and vertex positions and $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\mathbf{v}}$ their corresponding ground-truth values.

To precisely generate hand textures, we use a combination of loss functions. Given that the synthetic data were rendered using known ground-truth UV maps, we can directly enforce the model to produce textures that match the ground-truth UV maps with a UV loss:

$$\mathcal{L}_{uv} = \|UV^{\mathcal{G}} - UV^{0}\|_1 \tag{6}$$

where $UV^{\mathcal{G}}$ corresponds to the generated UV texture and UV^{O} to the ground-truth texture.

Additionally, we employ a differentiable renderer using an orthographic camera with trainable parameters that projects the generated 3D hand on the input image plane. To obtain accurate camera parameters and model the details of the appearance, we employ a pixel loss between the rendered image and the input image:

$$\mathcal{L}_{pix} = \|I^{\mathcal{R}} - I^{0}\|_1 \tag{7}$$

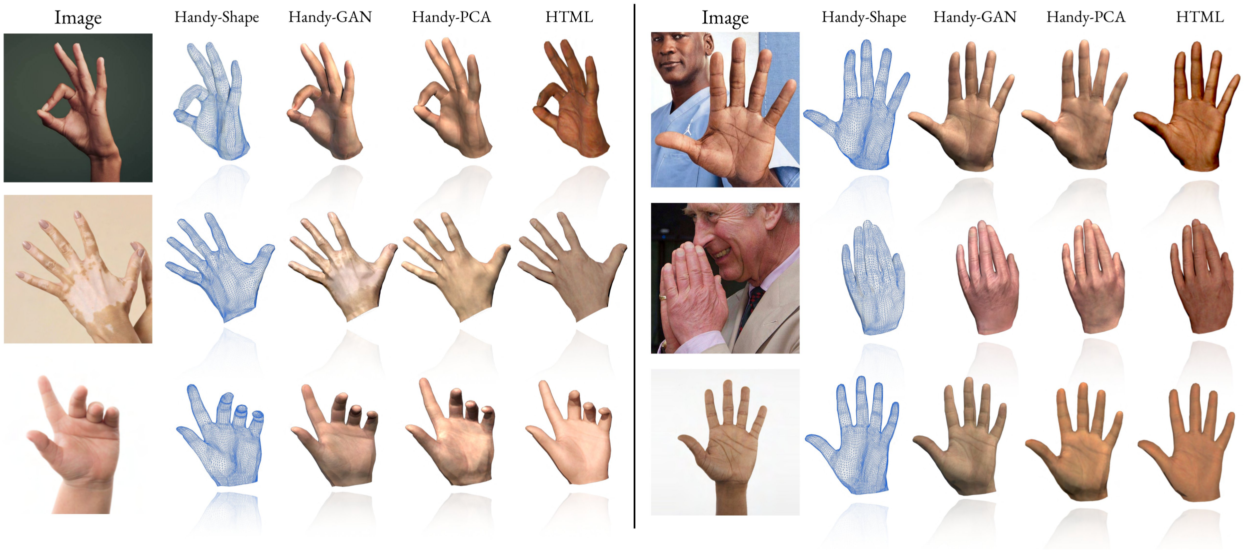

Figure 8. Hand shape and appearance reconstructions from single "in-the-wild" images. From left to right, we: i) depict the "in-the-wild" image, ii) our Handy-Shape reconstruction, iii) our Handy-GAN result, iv) our Handy-PCA model, and v) the HTML [37] texture on top of our shape mesh (Handy-Shape).

with $I^{\mathcal{R}}, I^0$ the original and the rendered images, respectively.

Finally, to constrain the generated hand textures, we apply a perceptual loss [52] that imposes the texture model to produce realistic textures that match the input image:

$$\mathcal{L}_{lpips} = \mathcal{F}(I^{\mathcal{R}}, I^0) \tag{8}$$

The exact network architecture is differed in the supplementary material.

Although synthetic data can be sufficient to train a hand pose and appearance estimation network, they usually constrain the texture regressor to latent codes that lie within the distribution of our textures, failing to reconstruct more challenging textures. In order to boost high fidelity appearance reconstruction, we collect a set of "in-the-wild" images and predict Handy pose, shape and texture parameters using the pre-trained regression network. Then, similar to [21], we only further optimize the texture parameters $\mathbf{w}$ to generate high resolution textures that match the appearance of the "in-the-wild" images. The optimization function is constructed with Eq. 7, 8, along with a L_2 regularization on $\mathbf{w}$ to make sure it does not greatly deviate from the initial estimation. Once we acquire the improved $\mathbf{w}'$, we fine-tune the regression network on the "in-the-wild" dataset.

To quantitatively assess the texture reconstruction of the proposed method, we feed the network with images from the scanning device used in the data. As ground-truth UV textures we use the corresponding UV maps of each subject acquired after the registration step. As can be seen in Table 2, the proposed model outperforms the HTML model by a significant margin, in terms of L_1 and LPIPS losses. The superiority of the proposed method can be also validated in Figure 7. To properly compare the texture reconstruction of each method, all three methods share the same shape and pose extracted from our regression network. The proposed method can reconstruct high frequency details of the input image such as wrinkles, rings, tattoos, and nail polish. In contrast, PCA-based methods produce smooth results that lack high frequency details and even fail to properly reconstruct the skin color (Figure 7, row 2).

Method	L_1 ($\times 10^{-2}$)	LPIPS [52]
HTML	2.14	0.092
Handy-PCA	1.44	0.065
Handy-GAN	**0.47**	**0.010**

Table 2. Quantitative comparison between the texture reconstruction models.

Additionally, we qualitatively compared the proposed method with HTML in an unconstrained setting using "in-the-wild" images. In Figure 8, we compare the three methods using challenging figures with different skin colors, shape structures, and light conditions. Similar to the previous experiment, all three methods share the same shape and pose. As can be easily seen, Handy-GAN can reconstruct high frequency details such as wrinkles and precise hand colors, even with hands that are out of the trained distribution. It is important to also note that Handy-GAN can also reconstruct textures from hands with vitiligo disorder that have severe color discontinuities.

Finally, to quantitatively evaluate shape and pose recon-

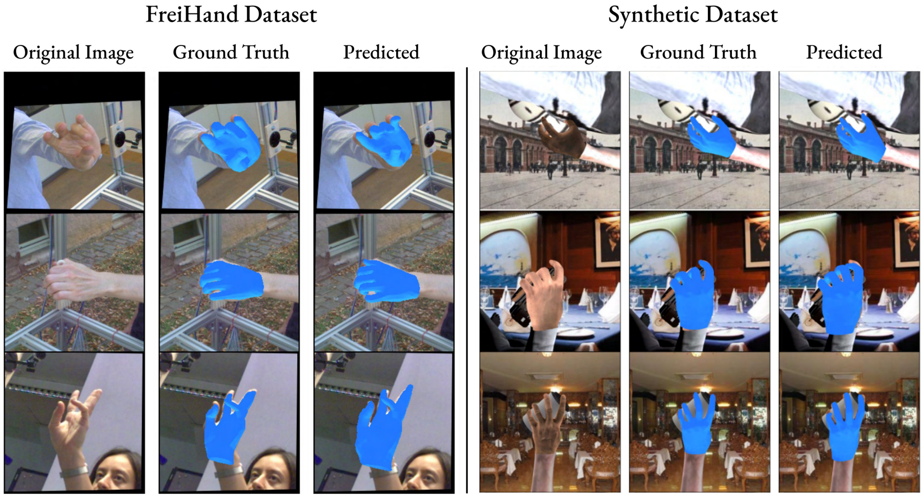

Figure 9. Shape and pose reconstructions from the FreiHand [55] and the proposed synthetic dataset.

Method	MPVPE ↓	MPJPE ↓	F@5 mm ↑	F@15 mm ↑
Hasson *et al.* [16]	13.2	-	0.436	0.908
Boukhayma *et al.* [3]	13.	-	0.435	0.898
MANO CNN [55]	10.8		0.529	0.935
MANO FIT [55]	13.7	-	0.439	0.892
HTML [37]	11.1	11.0	0.508	0.930
S^{2}Hand [6]	11.8	11.9	0.48	0.920
Ren *et al.* [38]	8.1	8.0	0.649	0.966
Proposed w/Obman	9.9	9.7	0.572	0.922
Proposed w/Synthetic	8.8	8.7	0.612	0.952
Proposed	**7.8**	**7.8**	**0.654**	**0.971**

Table 3. Quantitative comparison on the FreiHand dataset [55]

struction under "in-the-wild" conditions, we tested the performance of our model on the popular benchmark dataset FreiHand. Table 3 shows that the proposed method outperforms current state-of-the-art model-based methods utilizing MANO as their backbone. It is also important to note that, as expected, the proposed method trained on the proposed synthetic dataset (Proposed w/Synthetic), achieves better hand reconstructions compared to our method trained with the Obman dataset [16] (Proposed w/Obman). Such a finding validates our assumptions that the proposed synthetic dataset bridges the domain gap between synthetic and "in-the-wild" images. Qualitative results of our hand reconstruction are shown in Figure 9.

4.4. Reconstruction from Point Clouds

We also evaluated the proposed parametric model on shape and pose reconstruction from point clouds. In particular, we compared the proposed model with the state-of-the-art implicit hand model LISA [8] on the registered MANO dataset [39]. We follow [8] and sample $100K$ points on the surface of the MANO scans and measure the vertex-to-point distance (in mm) from the reconstruction to the scan (R2S) point cloud and the other way around (S2R). Table 4 shows that the proposed model achieves a lower reconstruction error with only 30 shape components, outperforming LISA and MANO models. Additionally, the proposed model using the MANO template and only 10 components outperforms the original MANO model by a large margin.

	R2S [mm]	S2R [mm]
MANO [39]	2.90	1.52
LISA-im [8]	1.96	1.13
LISA [8]	0.64	0.58
Proposed w/MANO (n_c = 10)	0.21	0.29
Proposed w/MANO (n_c = 30)	0.12	0.21
Proposed (n_c = 10)	0.16	0.25
Proposed (n_c = 30)	**0.11**	**0.19**

Table 4. Reconstruction error on point clouds sampled from the MANO dataset [39].

5. Conclusion

In this paper, we propose *Handy*, the first large-scale shape and appearance hand model that is composed of over 1200 subjects. Given the large demographic diversity of the subjects, the proposed model has more expressive power and overcomes the limitations of previous parametric models to reconstruct the shape of children's hands. Additionally, we train a style-based GAN to generate UV textures with high frequency details that traditional PCA methods fail to model. Under a series of experiments, we showcase the expressive power of Handy to reconstruct challenging hand shapes and appearances.

Acknowledgements: S. Zafeiriou and part of the research was funded by the EPSRC Fellowship DEFORM: Large Scale Shape Analysis of Deformable Models of Humans (EP/S010203/1).

References

[1] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 4

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 3

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 2, 3, 6, 8

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 3

[5] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20544–20554, 2022. 3

[6] 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. 8

[7] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European Conference on Computer Vision*, pages 231–248. Springer, 2022. 3

[8] 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. 2, 8

[9] Martin de La Gorce, David J Fleet, and Nikos Paragios. Model-based 3d hand pose estimation from monocular video. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1793–1805, 2011. 2, 3

[10] Martin de La Gorce, Nikos Paragios, and David J Fleet. Model-based hand tracking with texture, shading and self-occlusions. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[11] Nicolae Duta. A survey of biometric technology based on hand shape. *Pattern recognition*, 42(11):2797–2806, 2009. 1

[12] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10849–10859, 2021. 3

[13] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019. 3

[14] Lim Guan Ming, Jatesiktat Prayook, and Ang Wei Tech. Mobilehand: Real-time 3d hand shape and pose estimation from color image. In *27th International Conference on Neural Information Processing (ICONIP)*, 2020. 6

[15] Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. Tracking a hand manipulating an object. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1475–1482. IEEE, 2009. 3

[16] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2, 3, 5, 6, 8

[17] Tony Heap and David Hogg. Towards 3d hand tracking using a deformable model. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 140–145. Ieee, 1996. 3

[18] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 2, 3

[19] Linjun Jiang, Hailun Xia, and Caili Guo. A model-based system for real-time articulated hand tracking using a simple data glove and a depth camera. *Sensors*, 19(21):4680, 2019. 2

[20] 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. 4

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 7

[22] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 3

[23] Sameh Khamis, Jonathan Taylor, Jamie Shotton, Cem Keskin, Shahram Izadi, and Andrew Fitzgibbon. Learning an efficient model of hand shape variation from depth images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2540–2548, 2015. 3

[24] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020. 3

[25] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael M. Bronstein, and Stefanos Zafeiriou. Single image 3d hand reconstruction with mesh convolutions. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 3

[26] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 3

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3, 4

[28] Stan Melax, Leonid Keselman, and Sterling Orsten. Dynamics based 3d skeletal hand tracking. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 184–184, 2013. 3

[29] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 3

[30] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018. 3

[31] Evonne Ng, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11865–11874, 2021. 2

[32] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Training a feedback loop for hand pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3316–3324, 2015. 3

[33] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BmVC*, volume 1, page 3, 2011. 2

[34] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018. 3

[35] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 3

[36] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1106–1113, 2014. 2

[37] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. 2, 3, 4, 7, 8

[38] Jinwei Ren, Jianke Zhu, and Jialiang Zhang. End-to-end weakly-supervised multiple 3d hand mesh reconstruction from single image. *arXiv preprint arXiv:2204.08154*, 2022. 8

[39] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 2, 3, 4, 5, 8

[40] Raul Sanchez-Reillo, Carmen Sanchez-Avila, and Ana Gonzalez-Marcos. Biometric identification through hand geometry measurements. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1168–1171, 2000. 1

[41] Tanner Schmidt, Richard A Newcombe, and Dieter Fox. Dart: Dense articulated real-time tracking. In *Robotics: Science and Systems*, volume 2, pages 1–9. Berkeley, CA, 2014. 3

[42] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3633–3642, 2015. 2

[43] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 3

[44] Srinath Sridhar, Helge Rhodin, Hans-Peter Seidel, Antti Oulasvirta, and Christian Theobalt. Real-time hand tracking using a sum of anisotropic gaussians model. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 319–326. IEEE, 2014. 2

[45] Jonathan Taylor, Richard Stebbing, Varun Ramakrishna, Cem Keskin, Jamie Shotton, Shahram Izadi, Aaron Hertzmann, and Andrew Fitzgibbon. User-specific hand modeling from monocular depth sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 644–651, 2014. 3

[46] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM Transactions on Graphics (ToG)*, 35(6):1–11, 2016. 3

[47] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 2, 3

[48] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 3

[49] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2750–2760, 2022. 2, 3

[50] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3895–3905, 2022. 3

[51] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 2, 4

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[53] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 3

[54] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 3

[55] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 2, 6, 8