HandVoxNet: Deep Voxel-Based Network for 3D Hand Shape and Pose Estimation from a Single Depth Map

Jameel Malik[1,2,3] Ibrahim Abdelaziz[1,2] Ahmed Elhayek[2,4] Soshi Shimada[5]
Sk Aziz Ali[1,2] Vladislav Golyanik[5] Christian Theobalt[5] Didier Stricker[1,2]

[1]TU Kaiserslautern [2]DFKI Kaiserslautern [3]NUST Pakistan [4]UPM Saudi Arabia [5]MPII Saarland

Abstract

3D hand shape and pose estimation from a single depth map is a new and challenging computer vision problem with many applications. The state-of-the-art methods directly regress 3D hand meshes from 2D depth images via 2D convolutional neural networks, which leads to artefacts in the estimations due to perspective distortions in the images.

In contrast, we propose a novel architecture with 3D convolutions trained in a weakly-supervised manner. The input to our method is a 3D voxelized depth map, and we rely on two hand shape representations. The first one is the 3D voxelized grid of the shape which is accurate but does not preserve the mesh topology and the number of mesh vertices. The second representation is the 3D hand surface which is less accurate but does not suffer from the limitations of the first representation. We combine the advantages of these two representations by registering the hand surface to the voxelized hand shape. In the extensive experiments, the proposed approach improves over the state of the art by 47.8% on the SynHand5M dataset. Moreover, our augmentation policy for voxelized depth maps further enhances the accuracy of 3D hand pose estimation on real data. Our method produces visually more reasonable and realistic hand shapes on NYU and BigHand2.2M datasets compared to the existing approaches.

1. Introduction

The problem of deep learning-based 3D hand pose estimation has been extensively studied in the past few years [33], and recent works achieve high accuracy on public benchmarks [32, 18, 22]. Simultaneous estimation of 3D hand pose and shape from a single depth map is a newly emerging computer vision problem. It is more challenging than the pose estimation because annotating real images for shape is laborious and cumbersome. Other salient challenges include varying hand shapes, occlusions, high number of degrees of freedom (DOF) and self-similarity.

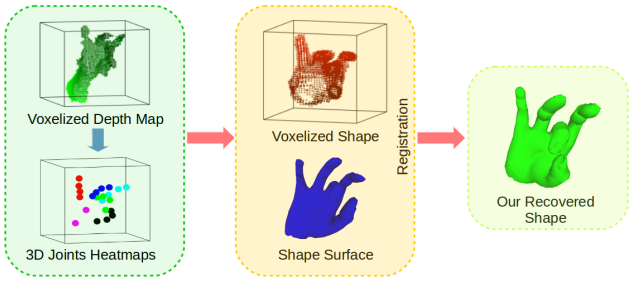


Figure 1: Hand shape and pose estimation with HandVoxNet. A 3D voxelized depth map and accurately regressed heatmaps of 3D joints (left block) are used to estimate two hand shape representations (middle block). To combine the advantages of these representations, we accurately register the shape surface to the voxelized shape (right block). Our architecture with 3D convolutions establishes a one-to-one mapping between voxelized depth map, voxelized hand shape and heatmaps of 3D joints.

The dense 3D hand mesh is a richer representation which is more useful than the sparse 3D joints, and it finds many applications in computer vision and graphics [27, 23, 17].

With the recent progress in deep learning, a few works [35, 7, 17, 13, 14] have proposed algorithms for simultaneous hand pose and shape estimation. Malik *et al.* [14] developed a 2D CNN-based approach that estimates shapes directly from 2D depth maps. The recovered shapes suffer from artifacts due to the limited representation capacity of their hand model [7, 17]. The same problem can occur even by embedding a realistic statistical hand model (*i.e.,* MANO) [23] inside a deep network [7, 35]. In contrast to these model-based approaches [35, 14], Ge *et al.* [7] proposed a more accurate direct regression-based approach using a monocular RGB image. Recently, Malik *et al.* [17] developed another direct regression-based approach from a single depth image. All of the approaches mentioned above treat and process depth maps with 2D CNNs, even though depth maps are intrinsically a 3D data. Training a 2D CNN to estimate 3D hand pose or shape given 2D representation of a depth map is highly non-linear and results in

perspective distortions in the estimated outputs [18]. V2V-PoseNet [18] is the first work that uses 3D voxelized grid of depth map to estimate 3D joints heatmaps and, thus, avoids perspective distortions. However, extending this work for shape estimation by directly regressing 3D heatmaps of mesh vertices is not feasible in practice.

In this work, we propose the first 3D CNN architecture which simultaneously estimates 3D shape and 3D pose given a voxelized depth map (see Fig. 1) To this end, we introduce novel architectures based on 3D convolutions which estimate two different representations of hand shape (Secs. 3–5). The first representation is the hand shape on a voxelized grid. It is estimated from a new *voxel-to-voxel* network which establishes a one-to-one mapping between the voxelized depth map and the voxelized shape. However, the estimated voxelized shape does not preserve the hand mesh topology and the number of vertices. For this reason, we also estimate hand surface (the second representation) with our *voxel-to-surface* network. Since this network does not establish a one-to-one mapping, the accuracy of the estimated hand surface is low but the hand topology is preserved. To combine the advantages of both representations, we propose registration methods to fit the hand surface to voxelized shape. Since real hand shape annotations are not available, we employ two 3D CNN-based synthesizers which act as sources of weak supervision by generating voxelized depth maps from our shape representations (see Fig. 2). To increase the robustness and accuracy of the hand pose estimation, we perform 3D data augmentation on the voxelized depth maps (Sec. 4.2).

We conduct ablation studies and perform extensive evaluations of the proposed method on real and synthetic datasets. Our approach improves the accuracy of hand shape estimation by 47.8% on SynHand5M dataset [14] and outperforms the state-of-the-art. Our method produces visually more reasonable and plausible hand shapes of NYU and BigHand2.2M datasets compared to the state-of-the-art approaches (Sec. 6). To summarise, our **contributions** are:

1. The first voxel-based hand shape and pose estimation approach with the following novel components:
   (i) *Voxel-to-voxel* 3D CNN-based network.
   (ii) *Voxel-to-surface* 3D CNN-based network.
   (iii) 3D CNN-based voxelized depth map synthesizers.
   (iv) Hand shape registration components.
2. A new 3D data augmentation policy on voxelized grids of depth maps.

2. Related Work

We now discuss the existing methods for deep hand pose and shape estimation. Moreover, we briefly report the most related works for depth-based hand pose estimation.

Deep Hand Pose and Shape Estimation. Malik *et al.* [14] proposed the first deep neural network for hand pose and shape estimation from a single depth image. To this end, they developed a model-based hand pose and shape layer which is embedded inside their deep network. Their approach suffers from artifacts due to the difficulty in optimizing complex hand shape parameters inside the network. Ge *et al.* [7] developed a direct regression-based algorithm for hand pose and shape estimation from a single RGB image. They highlight that the representation capacity of the statistical deformable hand model (*i.e.,* MANO [23]) could be limited due to the small amount of training data and the linear bases utilized for the shape recovery. Zhang *et al.* [35] introduced a similar MANO model based approach using a monocular RGB image. Recently, Malik *et al.* [17] proposed a structured weakly-supervised deep learning-based approach using a single depth image. All of the above-mentioned methods use 2D CNNs and treat the depth maps as 2D data. Consequently, the deep network is likely to produce perspective distortions in the shape and pose estimations [18]. In contrast, we propose the first 3D convolutions based architecture which establishes a one-to-one mapping between the voxelized depth map and the voxelized hand shape. This one-to-one mapping allows to more accurately reconstruct the hand shapes.

Hand Pose Estimation from Depth. In general, deep learning-based hand pose estimation methods can be classified into two categories. The first one encompasses the discriminative methods which directly estimate hand joint locations using CNNs [4, 3, 32, 21, 22, 18, 10, 6, 12]. The second category is hybrid methods which explicitly incorporate hand structure inside deep networks [16, 30, 8, 19, 15]. The disriminative methods achieve higher accuracy compared to the hybrid methods. The *voxel-to-voxel* approach [18] is powerful and highly effective because it uses 3D convolutions to learn a one-to-one mapping between the 3D voxelized depth map and 3D heatmaps of hand joints. Notably, the voxelized representation of depth maps is best suited for 3D data augmentation to improve the robustness and accuracy of the estimations. A few methods perform data augmentation on depth maps [19, 28] or voxelized depth maps [18]. In this work, we integrate the *voxel-to-voxel* approach with our pipeline and, additionally, perform new 3D data augmentation on voxelized depth maps. Our 3D data augmentation policy helps to achieve a noticeable improvement in the 3D pose estimation accuracy on real datasets.

3. Method Overview

Given a single input depth image, our goal is to estimate N 3D hand joint locations $\mathcal{J} \in \mathcal{R}^{3\times \mathrm{N}}$ (*i.e.,* 3D pose) and $\mathrm{K} = 1193$ 3D vertex locations $\mathcal{V} \in \mathcal{R}^{3\times \mathrm{K}}$ (*i.e.,* 3D shape).

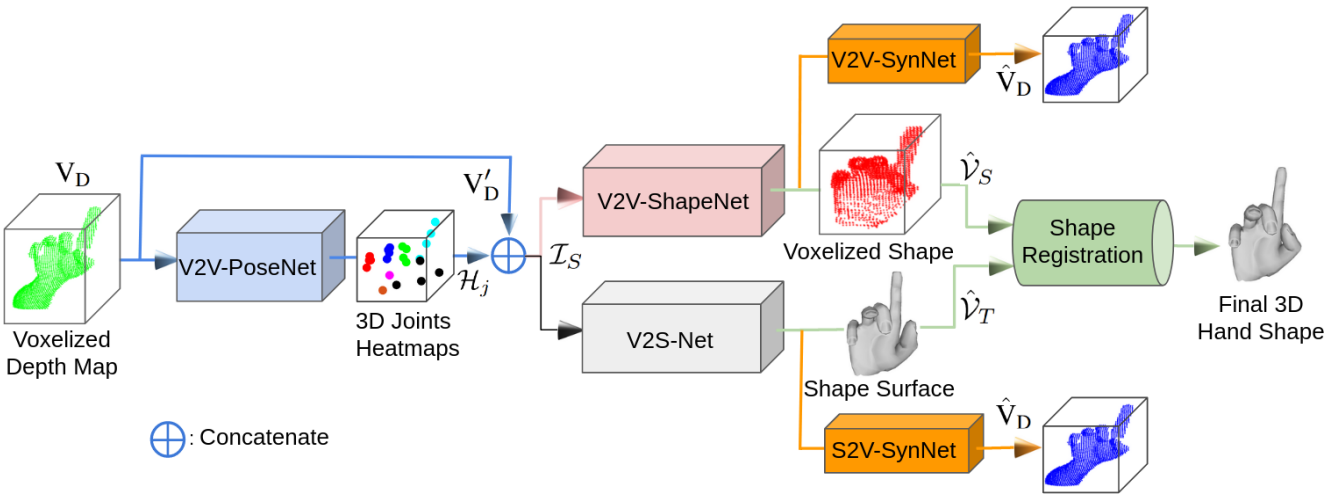


Figure 2: Overview of our approach for 3D hand shape and pose recovery from a 3D voxelized depth map. V2V-PoseNet estimates 3D joints heatmaps (*i.e.,* pose). Hand shape is obtained in two phases. First, V2V-ShapeNet and V2S-Net estimate the voxelized shape and shape surface, respectively. Thereby, V2V-SynNet and S2V-SynNet synthesize the voxelized depth acting as sources of weak-supervision. They are excluded during testing. In the second phase, shape registration accurately fits the shape surface to the voxelized shape.

Fig. 2 shows an overview of the proposed approach. The input depth image is converted into a voxelized grid (*i.e.,* $\mathrm{V_D}$) of size $88 \times 88 \times 88$, by using intrinsic camera parameters and a fixed cube size. For hand pose estimation, $\mathrm{V_D}$ is provided as an input to the *voxel-to-voxel* pose regression network (*i.e.,* V2V-PoseNet) that directly estimates 3D joint heatmaps $\{\mathcal{H}_j\}_{j=1}^{\mathrm{N}}$. Each 3D joint heatmap is represented as $44 \times 44 \times 44$ voxelized grid. We resize $\mathrm{V_D}$ to $44 \times 44 \times 44$ voxel grid size (*i.e.,* $\mathrm{V'_D}$) and concatenate it with the estimated $\mathcal{H}_j$, to provide as an input to our shape estimation network. We call this concatenated input as $\mathcal{I}_S$.

The voxelized hand shape (*i.e.,* $64 \times 64 \times 64$ grid size) is directly regressed via 3D CNN-based *voxel-to-voxel* shape regression network (*i.e.,* V2V-ShapeNet), by using $\mathcal{I}_S$ as an input. Notably, V2V-ShapeNet establishes a one-to-one mapping between the voxelized depth map and the voxelized shape. Therefore, it produces accurate voxelized shape representation but does not preserve the topology of hand mesh and the number of mesh vertices. To regress hand surface, $\mathcal{I}_S$ is fed to the 3D CNN-based *voxel-to-surface* regression network (*i.e.,* V2S-Net). Since the mapping between $\mathcal{I}_S$ and hand surface is not one-to-one, it is therefore less accurate. *Voxel-to-voxel* and *surface-to-voxel* synthesizers (*i.e.,* V2V-SynNet and S2V-SynNet) are connected after V2V-ShapeNet and V2S-Net, respectively. These synthesizers reconstruct $\mathrm{V'_D}$ and act as sources of weak supervision during training. They are excluded during testing. To combine the advantages of the two shape representations, we register the estimated hand surface to the estimated voxelized hand shape. We employ 3D CNN-based DispVoxNet [25] for synthetic data, and non-rigid gravitational approach (NRGA) [1] for real data.

4. The Proposed HandVoxNet Approach

In this section, we explain our proposed HandVoxNet approach by highlighting the function and effectiveness of each of its components. We develop an effective solution that produces reasonable hand shapes via 3D CNN-based deep networks. To this end, our approach fully exploits accurately estimated heatmaps of 3D joints as a strong pose prior, as well as voxelized depth maps. Given that collecting accurate real hand shape ground truth is hard and laborious, we develop a weakly-supervised network for real hand shape estimation by learning from accurately labeled synthetic data. Moreover, our 3D data augmentation on voxelized depth maps allows to further improve the accuracy and robustness of 3D hand pose estimation.

4.1. 3D Hand Shape Estimation

As aforementioned, estimating 3D hand shape from a 2D depth map by using 2D CNN is a highly non-linear mapping. It compels the network to perform perspective-distortion-invariant estimation which causes difficulty in learning the shapes. To address this limitation, we develop a full voxel-based deep network that effectively utilizes the estimated 3D pose and voxelized depth map to produce reasonable 3D hand shapes. Our proposed approach for 3D shape estimation comprises of two main phases. In the first phase, we estimate the shape surface and the voxelized hand shape. In the second phase, we register the estimated shape surface to the estimated voxelized hand shape by employing a 3D CNN-based registration for synthetic data and NRGA-based fitting process for real data.

Voxelized Shape Estimation. Our idea is to estimate 3D hand shape in the voxelized form via 3D CNN-based net-

work. It allows the network to estimate the shape in such a way that minimizes the chances for perspective distortion. Inspired by the approach proposed in the recent work [17], we consider sparse 3D joints as the latent representation of dense 3D shape. However, in this work, we combine 3D pose with the depth map which helps to represent the shape of hand more accurately. Furthermore, here we use more accurate and useful representations of 3D pose and 2D depth image which are 3D joints heatmaps and a voxelized depth map, respectively. The V2V-ShapeNet module is shown in Fig. 2. It can be considered as the 3D shape decoder:

$$\hat{\mathcal{V}}_S \sim \text{Dec}(\mathcal{H}_j \oplus \text{V}'_\text{D}) = p(\mathcal{V}_S|\mathcal{I}_S) \tag{1}$$

where $p(\mathcal{V}_S|\mathcal{I}_S)$ is the decoded distribution. The decoder learns to reconstruct the voxelized hand shape $\hat{\mathcal{V}}_S$ as close as possible to the ground truth voxelized hand shape $\mathcal{V}_S$. The V2V-ShapeNet is a 3D CNN-based architecture that directly estimates the probability of each voxel in the voxelized shape indicating whether it is the background (*i.e.,* 0) or the shape voxel (*i.e.,* 1). The per-voxel binary cross entropy loss $\mathcal{L}_{\mathcal{V}_S}$ for voxelized shape reconstruction reads:

$$\mathcal{L}_{\mathcal{V}_S} = -(\mathcal{V}_S \ \log(\hat{\mathcal{V}}_S) + (1 - \mathcal{V}_S) \ \log(1 - \hat{\mathcal{V}}_S)) \tag{2}$$

where $\mathcal{V}_S$ and $\hat{\mathcal{V}}_S$ are the ground truth and the estimated voxelized hand shapes, respectively. The architecture of V2V-ShapeNet is provided in the supplement.

Since the annotations for real hand shapes are not available, weak supervision is therefore essential in order to effectively learn real hand shapes. For this reason, we propose a 3D CNN-based V2V-SynNet (see Fig. 2) which acts as a source of weak supervision during training. This module is removed during testing. V2V-SynNet synthesizes the voxelized depth map from the estimated voxelized shape representation. The per-voxel binary cross entropy loss $\mathcal{L}^{\text{v}}_{\text{V}_\text{D}}$ for voxelized depth map reconstruction is given by:

$$\mathcal{L}^{\text{v}}_{\text{V}_\text{D}} = -(\text{V}_\text{D} \ \log(\hat{\text{V}}_\text{D}) + (1 - \text{V}_\text{D}) \ \log(1 - \log(\hat{\text{V}}_\text{D})) \tag{3}$$

where V_D and $\hat{\text{V}}_\text{D}$ are the ground truth and the reconstructed voxelized depth maps, respectively. The architecture of V2V-SynNet is provided in the supplement.

Shape Surface Estimation. The hand poses of the shape surfaces and voxelized shapes need to be similar for an improved shape registration. To facilitate the registration, we employ V2S-Net deep network which directly regresses $\mathcal{V}$. Based on the similar concept of hand shape decoding (as mentioned before), $\mathcal{I}_S$ is provided as an input to this network while the decoded output is the reconstructed hand mesh (see Fig. 2). The hand shape surface reconstruction loss $\mathcal{L}_{\mathcal{V}_T}$ is given by the standard Euclidean loss as:

$$\mathcal{L}_{\mathcal{V}_T} = \frac{1}{2} \left\| \hat{\mathcal{V}}_T - \mathcal{V}_T \right\|^2, \tag{4}$$

where $\mathcal{V}_T$ and $\hat{\mathcal{V}}_T$ are the respective ground truth and reconstructed hand shape surfaces. As explained before, in the case of missing real hand shape ground truth, the weak supervision on mesh vertices is provided by S2V-SynNet. In this case, the input to the S2V-SynNet is $\hat{\mathcal{V}}_T$ which is in 3D coordinates form. The loss function $\mathcal{L}^{\text{s}}_{\text{V}_\text{D}}$ for the S2V-SynNet is similar to Eq. (3). Further details on S2V-SynNet and V2S-Net can be found in the supplement.

CNN-based Shape Registration. Thanks to fully connected (FC) layers, V2S-Net is able to estimate hand shapes while preserving the order and number of points. Losing local spatial information is also known as a drawback of FC layers. In contrast to FC layers, a lot of works show fully convolutional networks (FCN) perform well in geometry regression tasks [24, 9, 18, 31]. However, estimating the voxelized hand shape by 3D convolutional layer results in an inconsistent number of points and loses point order. Hence, the ideal architecture is a network which estimates the hand shape without losing local spatial information while preserving the topology of the hand shape. To achieve this, we register the estimated shape by V2S-Net to the probabilistic shape representation estimated by FCN (V2V-ShapeNet) using DispVoxNets pipeline [25].

The original DispVoxNets pipeline is comprised of two stages, *i.e.,* global displacement estimation and refinement stage. The refinement stage is used to remove roughness on the point set surface. In contrast to the original approach, we replace the refinement stage with Laplacian smoothing [29]. This is possible because we assume the mesh topology is already known, and it is preserved by our pipeline.

In the DispVoxNet pipeline, the hand surface shape $\hat{\mathcal{V}}_T$ is first converted into a voxelized grid $\hat{\mathcal{V}}'_T$ (*i.e.,* $64 \times 64 \times 64$ voxelized grid size). DispVoxNet estimates per-voxel displacements of the dimension $64^3 \times 3$ between the reference $\hat{\mathcal{V}}_S$ and voxelized hand surface $\hat{\mathcal{V}}'_T$[1]. The displacement loss $\mathcal{L}_{\text{Disp}}$ is given by:

$$\mathcal{L}_{Disp.} = \frac{1}{Q^3} \left\| \mathbf{d} - D_{vn}(\hat{\mathcal{V}}_S, \hat{\mathcal{V}}'_T) \right\|^2, \tag{5}$$

where Q and $\mathbf{d}$ are the voxel grid size and the ground truth displacement, respectively. Since it is difficult to obtain $\mathbf{d}$ between the voxelized shape $\hat{\mathcal{V}}_S$ and hand surface $\hat{\mathcal{V}}_T$, the displacements are first computed between $\mathcal{V}_T$ and $\hat{\mathcal{V}}_T$, and are discretized to obtain $\mathbf{d}$. For more details of ground truth voxelized grid computation, please refer to [25].

NRGA-Based Shape Registration. In our voxel-based 3D hand shape and pose estimation pipeline (Fig. 2), the DispVoxNet [25] component requires shape annotations in its source-to-target displacement field learning phase. These annotations are available only for the synthetic dataset

[1] a larger grid size results in higher accuracy of DispVoxNet, and we hence set it to the maximum which our hardware supports.

which leaves a domain gap on the performance of DispVoxNet when tested on real dataset. To bridge this gap, we apply NRGA [1] to improve $\hat{\mathcal{V}}_T$ by registering it with $\hat{\mathcal{V}}_S$. NRGA is selected for this deformable alignment task over other methods [20, 2], as it supports local topology preservation of input hand surface and is robust at noise handling. Although NRGA is a point cloud alignment method, it provides the option to relax the deformation magnitude in the neighbouring regions of the hand mesh vertices. The original NRGA estimates a rigid transformation for every vertex $v \in \hat{\mathcal{V}}_T$ and diffuses the transformations in a subspace formed by a set of neighbourhood vertices of v. It builds a k-d tree on the template ($\hat{\mathcal{V}}_T$ in our case) and neighbourhood vertices are selected as the k-nearest neighbours (typically $0.1\% - 0.2\%$ of the total points in the template). We modify NRGA for the *surface-to-voxel* operation, *i.e.,* instead of k-nearest neighbours, we use connected vertices in a 4-ring for $\hat{\mathcal{V}}_T$. See more details in our supplementary material.

4.2. Data Augmentation in 3D

Our method for hand shape estimation relies on the accuracy of the estimated 3D pose. Therefore, the hand pose estimation method has to be accurate and robust. Training data augmentation helps to improve the performance of a deep network [19]. Existing methods for hand pose estimation [19, 28] use data augmentation in 2D. This is mainly because these methods treat depth maps as 2D data. The representation of the depth map in voxelized form makes it convenient to perform data augmentation in all three dimensions. In this paper, we propose a new 3D data augmentation policy which improves the accuracy and robustness of hand pose estimation (see Sec. 6.3).

During V2V-PoseNet training, we apply simultaneous rotations in all three axes (x,y,z) to each 3D coordinate (i, j, k) of $\mathrm{V_D}$ and $\mathcal{H}_j$ by using Euler transformations:

$$[\hat{i}, \hat{j}, \hat{k}]^{\mathrm{T}} = [\mathrm{Rot_x}(\theta_{\mathrm{x}})] \times [\mathrm{Rot_y}(\theta_{\mathrm{y}})] \times [\mathrm{Rot_z}(\theta_{\mathrm{z}})][i, j, k]^{\mathrm{T}}, \tag{6}$$

where $(\hat{i}, \hat{j}, \hat{k})$ is the transformed voxel coordinate. $\mathrm{Rot_x}(\theta_{\mathrm{x}})$, $\mathrm{Rot_y}(\theta_{\mathrm{y}})$ and $\mathrm{Rot_z}(\theta_{\mathrm{z}})$ are 3×3 rotation matrices around x, y and z axes. The values for θ_{x}, θ_{y} and θ_{z} are selected randomly in the ranges $[-40°, +40°]$, $[-40°, +40°]$ and $[-120°, +120°]$, respectively. In addition to rotations in 3D, following [18], we perform scaling and translation in the respective ranges $[+0.8, +1.2]$ and $[-8, +8]$.

5. The Network Training

$\mathrm{V_D}$ is generated by projecting the raw depth image pixels into 3D space. Hand region points are then extracted by using a cube of size 300 that is centered on hand palm center position. 3D point coordinates of the hand region are discretized in the range $[1, 88]$. Finally, to obtain $\mathrm{V_D}$, the voxel value is set to 1 for the 3D point coordinate of hand region, and 0 otherwise. Following [18], $\mathcal{H}_j$ are generated as 3D Gaussians. Similar to the generating of $\mathrm{V_D}$, $\mathcal{V}_S$ is obtained by voxelizing the hand mesh. $\mathcal{V}_T$ is created by normalizing the mesh vertices in the range $[-1, +1]$. We perform this normalization by subtracting the vertices from the palm center and then dividing them by half of the cube size.

Methods	3D $\mathcal{V}$ Err. (*mm*)
V2S-Net (w/o $\mathcal{H}_j$)	8.78
V2S-Net (w/o $\mathrm{V'_D}$)	3.54
V2S-Net (with $\mathcal{H}_j \oplus \mathrm{V'_D}$)	**3.36**
Methods	3D $\mathcal{S}$ Err.
V2V-ShapeNet (w/o $\mathcal{H}_j$)	0.007
V2V-ShapeNet (w/o $\mathrm{V'_D}$)	0.016
V2V-ShapeNet (with $\mathcal{H}_j \oplus \mathrm{V'_D}$)	**0.005**

Table 1: Ablation study on inputs (*i.e.,* $\mathcal{H}_j$ and $\mathrm{V'_D}$) to V2S-Net and V2V-ShapeNet. We observe that combining both inputs is useful for these two networks.

We train V2V-PoseNet [18] on NYU [28], BigHand2.2M [34] and SynHand5M [14] datasets separately with the 3D data augmentation technique mentioned in Sec. 4.2. For SynHand5M dataset, we train V2S-Net and V2V-ShapeNet (including the synthesizers S2V-SynNet and V2V-SynNet) separately using RMSProp as an optimization method with a batch size of 8 and a learning rate LR = 2.5×10^{-4}. After training the pose and shape networks, we put these networks together in the pipeline (see Fig. 2) and fine-tune them in an end-to-end manner with synthetic, as well as combined real and synthetic data. The total loss $\mathcal{L}_{\mathrm{T}}$ read as follows:

$$\mathcal{L}_{\mathrm{T}} = \mathcal{L}_{\mathcal{H}} + \mathbb{1}\mathcal{L}_{\mathcal{V}_S} + \mathbb{1}\mathcal{L}_{\mathcal{V}_T} + \mathcal{L}^{\mathrm{v}}_{\mathrm{V_D}} + \mathcal{L}^{\mathrm{s}}_{\mathrm{V_D}} \tag{7}$$

where $\mathcal{L}_{\mathcal{H}}$ is heatmaps loss [18] and $\mathbb{1}$ represents an indicator function layer. This layer forwards the estimations to the loss layer only for synthetic data using a flag value, which is 1 for synthetic and 0 for real data. It disables the gradients flow during the backward pass in the case of real data. For fine-tunings, we use RMSProp with a batch size of 6 and a learning rate 2.5×10^{-5}. DispVoxNet is trained only on SynHand5M dataset due to the availability of the ground truth geometry. During the training, Adam optimizer [11] with a learning rate of 3.0×10^{-4} was employed. The training continues until the convergence of $\mathcal{L}_{\mathrm{Disp}}$ with batch size 12. All models are trained until convergence on a desktop workstation equipped with Nvidia Titan X GPU.

6. Experiments

We perform qualitative and quantitative evaluations of our complete pipeline including ablation studies on the fully labeled SynHand5M [14] dataset. We qualitatively evaluate

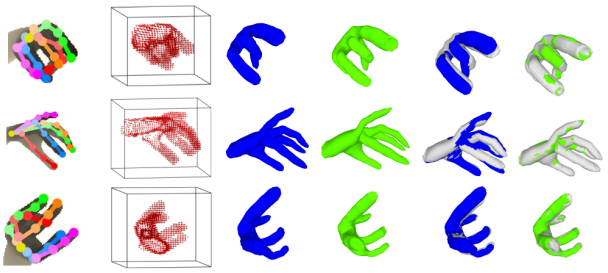


Figure 3: Qualitative results on SynHand5M [14] dataset. Estimated hand pose overlay (1st col), voxelized shape (2nd col), hand surface (3rd col), final shape (4th col), and the overlays of hand surface and final shapes with ground truth (gray color) are illustrated.

Methods	3D $\mathcal{V}$ Err. (*mm*)
DeepHPS [14]	11.8
WHSP-Net [17]	5.12
ours (w/o synthesizers)	2.92
ours (with synthesizers)	**2.67**

Table 2: Comparison with the state of the arts on SynHand5M [14]. Our full method, with V2V-SynNet and S2V-SynNet synthesizers, outperforms the WHSP-Net approach [17] by 47.85%.

real hand shape recovery on NYU [28] and BigHand2.2M [34] datasets. Furthermore, we study the impact of our 3D data augmentation on V2V-PoseNet [18].

6.1. Datasets and Evaluation Metrics

Although there are many depth-based hand pose datasets [34, 5, 26], only a few of them (*i.e.,* BigHand2.2M [34], NYU [28], SynHand5M [14]) provide adequate training data and annotations which resemble the joint locations of a real hand. NYU real benchmark offers joint annotations for 72757 and 8252 RGBD images of the training ($\mathcal{T}_{\text{N}}$) and test sets, respectively. Their hand model contains 42 DOF which makes it possible to combine this dataset with the recent benchmarks (*e.g.,* BigHand2.2M). BigHand2.2M is a million-scale real benchmark. For pose estimation, it provides accurate joint annotations for 956k training ($\mathcal{T}_{\text{B}}$) depth images acquired from 10 subjects. Their hand model contains 21 joint locations which resembles real hand skeleton. The size of the BigHand2.2M's test set is 296k. The annotation of hand palm center is not given in the BigHand2.2M dataset. Hence, we obtain the hand palm center position by taking the average of the metacarpal joints and the wrist joint positions. SynHand5M dataset contains fully annotated 5 million depth images for both the 3D hand pose and shape. The sizes of its training ($\mathcal{T}_{\text{S}}$) and test sets are 4.5M and 500k, respectively. The joint annotations of BigHand2.2M are fully compatible with SynHand5M.

Components	runtime, *sec*
V2V-PoseNet	0.011
V2V-ShapeNet	0.0015
V2S-Net	0.0038
DispVoxNet (GPU + CPU)*	0.162
NRGA (CPU)	59 - 70

Table 3: Runtime: (first four rows) forward-pass of deep networks on GPU. "*" shows that Laplacian smoothing runs on CPU.

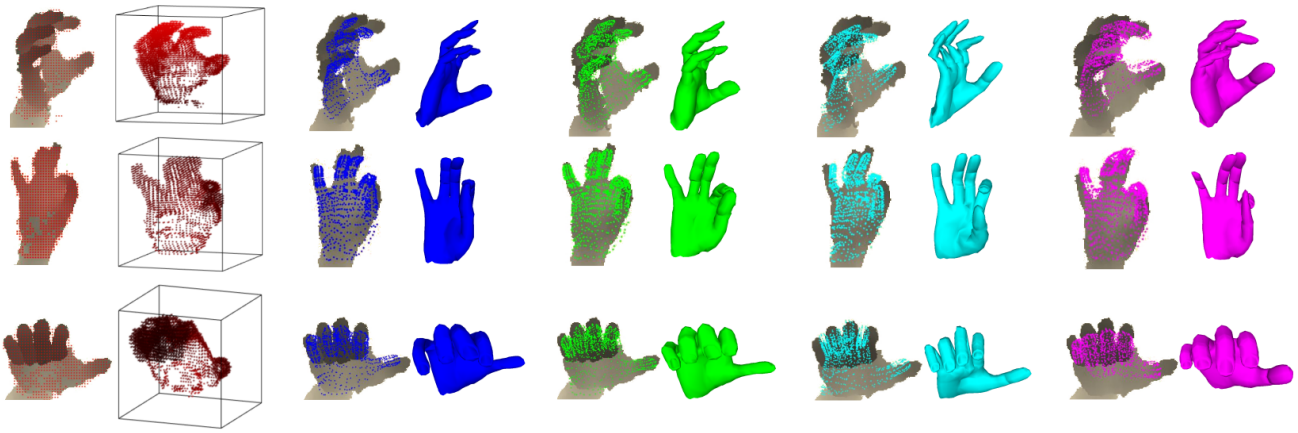
We use three evaluation metrics: (i) the average 3D joint location error over all test frames (3D $\mathcal{J}$ Err.); (ii) mean vertex location error over all test frames (3D $\mathcal{V}$ Err.); and (iii) mean voxelized shape error (*i.e.,* per-voxel binary cross entropy) over all test data (3D $\mathcal{S}$ Err.).

6.2. Evaluation of Hand Shape Estimation

In this subsection, we evaluate our method on SynHand5M, NYU and BigHand2.2M benchmarks.

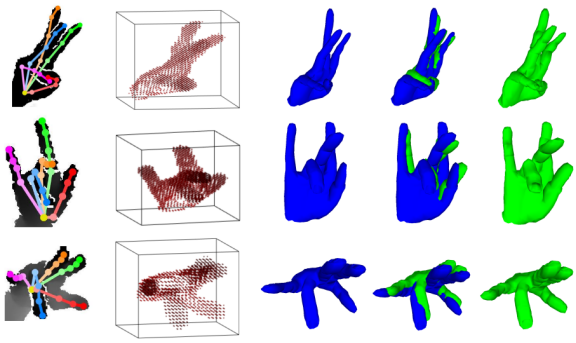
Synthetic Hand Shape Reconstruction. We train our complete pipeline on the fully labeled SynHand5M dataset by following the training methodology explained in Sec. 5. We conduct two ablation studies to show the effectiveness of our design choice. First is the regression of $\mathcal{V}_T$ and $\mathcal{V}_S$ by using input V'_{D} (*i.e.,* without $\mathcal{H}_j$) and the synthesizers. Similar experiments are repeated by using $\mathcal{H}_j$ (*i.e.,* without V'_{D}) and $\mathcal{I}_S$ (*i.e.,* with $\mathcal{I}_S \oplus \text{V}'_{\text{D}}$) as separate inputs to V2V-ShapeNet and V2S-Net. The results are summarized in Table 1 and clearly show the benefit of concatenating voxelized depth map with 3D heatmaps. The second ablation study is to observe the impact of V2V-SynNet and S2V-SynNet, given $\mathcal{I}_S$ as an input to the complete shape estimation network. We train V2S-Net and V2V-ShapeNet with and without using their respective synthesizers (see Fig. 2). The quantitative results and comparisons with the state-of-the-art methods on SynHand5M test set are summarized in Table 2. Our method with synthesizers improves on ours without synthesizers, and achieves 47.8% improvement in the accuracy compared to the recent WHSP-Net [17]. Several synthesized samples of voxelized depth maps are shown in the supplement. The qualitative results of shape representations and poses are shown in Fig. 3. DispVoxNet fits the estimated hand surface to the estimated voxelized hand shape, thereby improving the hand surface reconstruction accuracy by 20.5% (*i.e.,* from $3.36mm$ to $2.67mm$). Notably, the accuracy of our hand surface estimation is higher compared to WHSP-Net (*cf.* Tables 1 and 2), which clearly shows the effectiveness of employing 3D CNN based network for mesh vertex regression.

Real Hand Shape Reconstruction. To estimate plausible real hand shape representations, the synthesizers are essential (see Fig. 2). For NYU hand surface and voxelized shape



(a) Voxelized shape (b) Shape surface (c) Ours (final shape) (d) DeepHPS [14] (e) WHSP-Net [17]

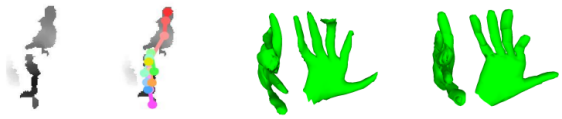
Figure 4: Shape reconstruction of NYU [28] dataset: (a), (b) and (c) show the 2D overlays and 3D visualizations of estimated voxelized hand shape, shape surface, and the final shape after registration, respectively. (d) and (e) show the corresponding results of hand shapes from DeepHPS [14] and WHSP-Net [17] methods. Our approach produces visually more accurate hand shapes than the existing approaches.



(a) (b) (c) (d) (e)

Figure 5: Shape reconstruction of BigHand2.2M [34] dataset: (a) the 2D pose overlay; (b), (c) recovered voxelized shape and shape surface, respectively; (d) the overlays of shape surface and registered shape; (e) the final hand shape.

recovery, we combine the training sets of NYU and SynHand5M (*i.e.,* $\mathbf{T}_{\text{NS}} = \mathcal{T}_{\text{N}} + \mathcal{T}_{\text{S}}$) by selecting closely matching 22 common joint positions in both datasets. However, note that the common joint positions are still not exactly similar in both the datasets. V2S-Net and V2V-ShapeNet recover plausible hand shape representations while NRGA-based method performs a successful registration (as shown in Fig. 4(a), (b) and (c)). It is observed that the voxelized shape is more accurately estimated than the hand surface. Thereby, the alignment further refines the hand surface. Using the similar training strategy, we combine BigHand2.2M and SynHand5M datasets and shuffle them (*i.e.,* $\mathbf{T}_{\text{BS}} = \mathcal{T}_{\text{B}} + \mathcal{T}_{\text{S}}$). Samples of the estimated hand shape representations for BigHand2.2M are shown in Fig. 5.



(a) Depth (b) 3D Pose (c) DispVoxNet [25] (d) NRGA [1]

Figure 6: Failure case: our method is unable to produce plausible shapes in cases of severe occlusion and missing depth information.

Methods	3D $\mathcal{J}$ Err. (*mm*)
DeepHPS [14]	6.30
WHSP-Net [17]	4.32
V2V-PoseNet [18]	3.81
our HandVoxNet (full method)	**3.75**

Table 4: 3D hand pose estimation results on SynHand5M [14] dataset. We compare the accuracy of our full method (*i.e.,* HandVoxNet) with state-of-the-art methods.

We qualitatively compare our reconstructed hand shapes of NYU dataset with the state of the art. For better illustration of the shape reconstruction accuracy, we show the 2D overlay of hand mesh onto the corresponding depth image (as shown in Fig. 4-(d) and (e)). Model-based DeepHPS [14] suffers from artifacts, the regression-based WHSP-Net approach [17] produces perspective distortions and incorrect sizes of shapes. In contrast, HandVoxNet recovers visually more plausible hand shapes (Fig. 4-(c)). Table 3 provides the runtimes of different components of our pipeline.

Failure Cases. Our approach fails to estimate plausible hand shapes in cases of severe occlusion of hand parts and

Dataset	Method	3D $\mathcal{J}$ Err. (*mm*)
NYU	V2V-PoseNet [18]	9.22
	V2V-PoseNet (our 3D augm.)	**8.72**
BigHand2.2M	V2V-PoseNet [18]	9.95
	V2V-PoseNet (our 3D augm.)	**9.27**

Table 5: 3D hand pose estimation results on NYU [28] and BigHand2.2M [34] datasets using our 3D data augmentation.

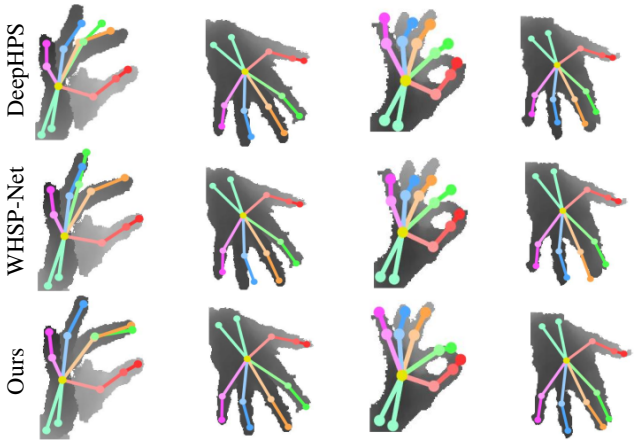


Figure 7: Samples of NYU [28] depth images with 2D overlay of the estimated 3D hand pose. Our method produces more accurate results compared to WHSP-Net [17] and DeepHPS [14] methods.

missing information in the depth map (see Fig. 6).

6.3. Evaluation of Hand Pose Estimation

In our approach, the accuracy of the estimated hand shape is dependent on the accuracy of estimated 3D pose (see Sec. 4). Therefore, the hand pose estimation needs to be robust and accurate. Therefore, we perform a new 3D data augmentation on voxelized depth maps which further improves the accuracy of 3D hand pose estimation on real datasets. Notably, our focus is to develop an effective approach for simultaneous hand pose and shape estimation. However, for completeness, we show our results and comparisons of hand pose estimation with SynHand5M, NYU and BigHand2.2M datasets.

SynHand5M dataset: We do not perform training data augmentation on SynHand5M because this dataset originally contains large viewpoint variations [14]. We train our full method and V2V-PoseNet [18] on SynHand5M dataset. The quantitative results on the test set are presented in Table 4. We observe that the backpropagation from the shape regression pipeline is effective and improves the accuracy of the estimated 3D pose. We achieve 13.19% improvement in the accuracy compared to WHSP-Net approach [17].

NYU and BigHand2.2M datasets: V2V-PoseNet [18] is a powerful pose estimation method that exploits the 3D data representations of hand pose and depth map. Thanks to our 3D data augmentation strategy (see Sec. 4.2), we improve

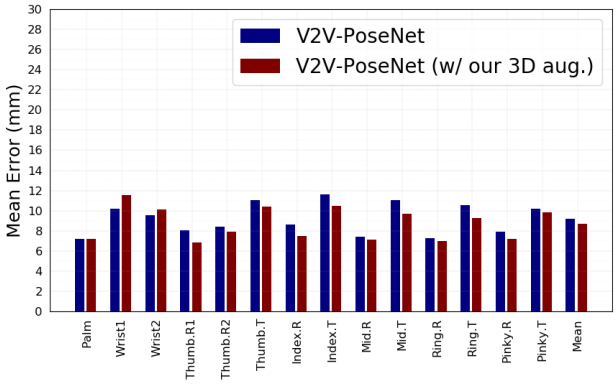


Figure 8: We study the impact of our 3D data augmentation on the pose estimation accuracy of V2V-PoseNet [18] on NYU [28] dataset. The graph shows mean errors for individual hand joints.

the accuracy by 5.42% and 6.83% compared to the original V2V-PoseNet models on NYU and BigHand2.2M datasets, respectively (see Table 5). Fig. 8 shows the average errors on individual hand joints. We observe a noticeable improvement in the accuracy of the finger tips. The qualitative results and comparisons with the state-of-the-art methods for hand pose estimation are shown in Fig. 7.

7. Conclusion and Future Work

We develop the first voxel-based pipeline for 3D hand shape and pose recovery from a single depth map, which establishes an effective inter-link between hand pose and shape estimations using 3D convolutions. This inter-link boosts the accuracy of both estimates, which is demonstrated experimentally. We employ 3D voxelized depth map and accurately estimated 3D heatmaps of joints as inputs to reconstruct two hand shape representations, *i.e.,* 3D voxelized shape and 3D shape surface. To combine the advantages of both shape representations, we employ registration methods, *i.e.,* DispVoxNet and NRGA, which accurately fit the shape surface to the voxelized shape.

The experimental evaluation further shows that our 3D data augmentation policy on voxelized grids enhances the accuracy of 3D hand pose estimation on real data. HandVoxNet produces visually more accurate hand shapes of real images compared to the previous methods. All these results indicate that the one-to-one mapping between voxelized depth map, voxelized shape and 3D heatmaps of joints is essential for an accurate hand shape and pose recovery.

In future work, generating a realistic synthetic dataset can further enhance the hand shape reconstruction from real images. The runtimes of the used registration methods can be improved by the parallelization on GPUs.

Acknowledgement: This work was funded by the German Federal Ministry of Education and Research as part of the project VIDETE (grant number 01IW18002) and the ERC Consolidator Grant 770784.

References

[1] Sk Aziz Ali, Vladislav Golyanik, and Didier Stricker. Nrga: Gravitational approach for non-rigid point set registration. In *International Conference on 3D Vision (3DV)*, 2018.

[2] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *Computer Vision and Pattern Recognition (CVPR)*, 2007.

[3] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *International Conference on Computer Vision (ICCV)*, 2019.

[4] Yujin Chen, Zhigang Tu, Liuhao Ge, Dejun Zhang, Ruizhi Chen, and Junsong Yuan. So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning. In *International Conference on Computer Vision (ICCV)*, 2019.

[5] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] Liuhao Ge, Zhou Ren, and Junsong Yuan. Point-to-point regression pointnet for 3d hand pose estimation. In *European Conference on Computer Vision (ECCV)*, 2018.

[9] Vladislav Golyanik, Soshi Shimada, Kiran Varanasi, and Didier Stricker. Hdm-net: Monocular non-rigid 3d reconstruction with learned deformation model. In *International Conference on Virtual Reality and Augmented Reality (EuroVR)*, 2018.

[10] Hengkai Guo, Guijin Wang, Xinghao Chen, Cairong Zhang, Fei Qiao, and Huazhong Yang. Region ensemble network: Improving convolutional network for hand pose estimation. In *International Conference on Image Processing (ICIP)*, 2017.

[11] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[12] Jameel Malik, Ahmed Elhayek, Sheraz Ahmed, Faisal Shafait, Muhammad Malik, and Didier Stricker. 3dairsig: A framework for enabling in-air signatures using a multi-modal depth sensor. *Sensors*, 18(11):3872, 2018.

[13] Jameel Malik, Ahmed Elhayek, Fabrizio Nunnari, and Didier Stricker. Simple and effective deep hand shape and pose regression from a single depth image. *Computers & Graphics*, 85:85–91, 2019.

[14] Jameel Malik, Ahmed Elhayek, Fabrizio Nunnari, Kiran Varanasi, Kiarash Tamaddon, Alexis Heloir, and Didier Stricker. Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. In *International Conference on 3D Vision (3DV)*, 2018.

[15] Jameel Malik, Ahmed Elhayek, and Didier Stricker. Simultaneous hand pose and skeleton bone-lengths estimation from a single depth image. In *International Conference on 3D Vision (3DV)*, 2017.

[16] Jameel Malik, Ahmed Elhayek, and Didier Stricker. Structure-aware 3d hand pose regression from a single depth image. In *International Conference on Virtual Reality and Augmented Reality (EuroVR)*, 2018.

[17] Jameel Malik, Ahmed Elhayek, and Didier Stricker. Whspnet: A weakly-supervised approach for 3d hand shape and pose recovery from a single depth image. *Sensors*, 19(17):3784, 2019.

[18] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[19] Markus Oberweger and Vincent Lepetit. Deepprior++: Improving fast and accurate 3d hand pose estimation. In *International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[20] Chavdar Papazov and Darius Burschka. Deformable 3d shape registration based on local similarity transforms. *Computer Graphics Forum*, pages 1493–1502, 2011.

[21] Georg Poier, Michael Opitz, David Schinagl, and Horst Bischof. Murauer: Mapping unlabeled real data for label austerity. In *Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[22] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Feature mapping for learning fast and accurate 3d pose inference from synthetic images. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):245, 2017.

[24] Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Didier Stricker. Ismo-gan: Adversarial learning for monocular non-rigid 3d reconstruction. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[25] Soshi Shimada, Vladislav Golyanik, Edgar Tretschk, Didier Stricker, and Christian Theobalt. Dispvoxnets: Non-rigid point set alignment with supervised learning proxies. In *International Conference on 3D Vision (3DV)*, 2019.

[26] James S Supancic, Grégory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan. Depth-based hand pose estimation: data, methods, and challenges. In *International Conference on Computer Vision (ICCV)*, 2015.

[27] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (TOG)*, 35(4):143, 2016.

[28] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands

using convolutional networks. *ACM Transactions on Graphics (TOG)*, 33(5):169, 2014.

[29] Jörg Vollmer, Robert Mencl, and Heinrich Mueller. Improved laplacian smoothing of noisy surface meshes. In *Computer Graphics Forum*, pages 131–138, 1999.

[30] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[31] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Neural Information Processing Systems (NeurIPS)*, 2016.

[32] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *International Conference on Computer Vision (ICCV)*, 2019.

[33] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] Shanxin Yuan, Qi Ye, Björn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[35] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *International Conference on Computer Vision (ICCV)*, 2019.