I2UV-HandNet: Image-to-UV Prediction Network for Accurate and High-fidelity 3D Hand Mesh Modeling

Ping Chen[1] Yujin Chen[2,3] Dong Yang[1] Fangyin Wu[1] Qin Li[1] Qingpei Xia[1] Yong Tan[1]
[1]IQIYI Inc. [2]Wuhan University [3]Technical University of Munich
{redcping, terencecyj}@gmail.com {yangdong01, wufangying, liqin01, xiaqingpei, tanyong}@qiyi.com

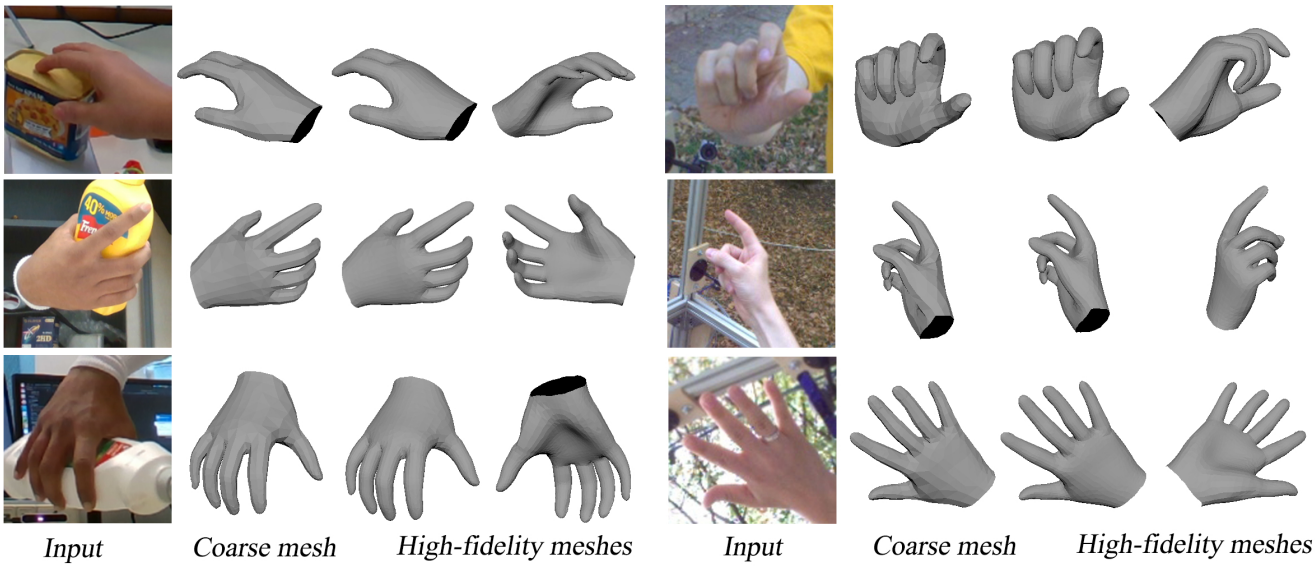

Figure 1: We propose a novel Image-to-UV prediction network (I2UV-HandNet) that estimates accurate and high-fidelity hand mesh from a single RGB image. Here, we present example results on the HO-3D dataset (left) and the FreiHAND dataset (right). From left to right: input, estimated coarse mesh, estimated high-fidelity mesh (in two viewpoints).

Abstract

Reconstructing a high-precision and high-fidelity 3D human hand from a color image plays a central role in replicating a realistic virtual hand in human-computer interaction and virtual reality applications. The results of current methods are lacking in accuracy and fidelity due to various hand poses and severe occlusions. In this study, we propose an I2UV-HandNet model for accurate hand pose and shape estimation as well as 3D hand super-resolution reconstruction. Specifically, we present the first UV-based 3D hand shape representation. To recover a 3D hand mesh from an RGB image, we design an AffineNet to predict a UV position map from the input in an image-to-image translation fashion. To obtain a higher fidelity shape, we exploit an additional SRNet to transform the low-resolution UV map outputted by AffineNet into a high-resolution one. For the first time, we demonstrate the characterization capability of the UV-based hand shape representation. Our experiments show that the proposed method achieves state-of-the-art performance on several challenging benchmarks.

1. Introduction

Observing and understanding the human hand has been an important task in computer vision and human-computer interaction, with applications from gesture recognition to augmented reality (AR) and virtual reality (VR). Recently, we have witnessed significant progress in 3D hand pose and shape estimation [5, 7, 8, 10, 14, 36, 52, 49], driven by efforts in large-scale data collection and annotation [46, 52, 53], coupled with the development of 3D representations and learning methods [6, 33]. This has led to remarkable advances in 3D hand understanding from a single-view color image.

Due to the lack of hand surface data, most of the earlier works study 3D pose estimation by estimating 3D joint location from a single image [3, 7, 20, 37, 52]. However, the sparse joints representation cannot meet the needs of

many applications such as interacting a virtual hand with an object in some immersive VR scenarios [19]. To better display the hand surface, previous approaches regress a parametric hand model (MANO) [34] with articulated and nonrigid deformations [5, 15, 26, 31, 48]. Although it is easy to use CNN to predict the MANO parameters from RGB input and use 3D annotations to supervise this regression process [17, 53], this high-dimensional nonlinear regression limits the accuracy of reconstructed hands. Then regression-based methods introduce various intermediate representations to guide the training process. In these methods, the 3D hand reconstruction is decomposed into two stages, that first regresses a set of intermediate representations such as 2D keypoints, masks, or 3D keypoints, then predicts the model parameters from these intermediate representations [48, 50]. The performance of these works largely depends on the design of these intermediate representations as well as the usage of reasonable supervision terms. More recently, [14, 27] remove the dependence of parametric model prior and directly regress the 3D coordinates of mesh vertices. Even though good performances are shown, the above methods, which estimate model parameters or vertices coordinate from high-dimensional encoded features, break the spatial relationship contained in the original pixel space. [29] proposes to predict a 1D heatmap for each mesh vertex coordinate and achieves state-of-the-art performance, but it only preserves spatial information in feature transformation while its vertex-wise output is still a discrete 3D representation. Different to above 3D representation and learning method, we propose to use UV position map [13] as the hand representation in this work.

Inspired by recent 3D body recovery methods that map a 3D mesh of the human body into a UV map representation [1, 43], we propose to represent 3D hand surface in UV space and train a neural network to predict 3D hand shape from a single RGB input. The usage of UV representation enables an efficient network to directly regress the hand surface, without relying on any model prior or intermediate representations. To properly predict the UV position map from the RGB input, we present AffineNet that addresses the single-view 3D reconstruction issue in an image-to-image translation task. Traditional image-to-image conversion pipelines are designed for tasks (such as appearance conversion or semantic segmentation) with good spatial alignment between the input and the output [41, 51]. However, in our setup, the hand shape displayed by the UV position map is different from that in the input RGB image. To address this problem, we propose a novel affine connection module to align the encoded feature maps with the UV maps and then connect the aligned feature maps with the decoded feature maps. In AffineNet, hierarchical UV position maps and multi-level feature maps are employed, and multiple UV maps can be supervised at the training stage. For 3D pose estimation, we obtain a set of 3D keypoints from the output hand mesh via a pre-defined mapping.

Another advantage of the UV-based representation is that the dense UV position map enables reconstructing a 3D surface with more vertices by sampling in the valid area of the UV position map. Motivated by this observation, we present a UV-based 3D hand super-resolution reconstruction module named SRNet to realize high-fidelity 3D hand reconstruction from the coarse 3D hand shape. In order to make the best of the proposed hand UV position map representation, we restore high-fidelity hand shape by using a CNN to map the low-resolution UV position map into a high-resolution one. However, there lacks of high-fidelity hand surface data to supervise the learning of SRNet. Thus, we construct a scan dataset called SuperHandScan to learn the SRNet. We transfer the high-quality 3D hand scan and the registered coarse MANO model to high/low-resolution UV position maps, and then use those UV maps to train the SRNet. Since the input of SRNet is a coarse hand mesh in UV-based representation, there is wide application scope for the SRNet, in other words, a well-trained SRNet can be used for mesh super-resolution reconstruction of any coarse hand mesh.

In summary, we present an I2UV-HandNet model which consists of an AffineNet for 3D hand pose and shape estimation and an SRNet for hand mesh super-resolution reconstruction. Overall, the main contributions of this paper are summarized as follows:

- To our best knowledge, we are the first to introduce UV map representation in 3D hand pose and shape estimation. Based on our novel representation, we propose an end-to-end network named AffineNet to predict hand mesh from a single color image.
- For the first time in hand reconstruction, we propose SRNet, a hand mesh super-resolution reconstruction network to predict a high-fidelity hand mesh from a coarse hand mesh.
- Our method can predict accurate and high-fidelity hand meshes from RGB inputs. Experimental results show that our method surpasses other state-of-the-art methods on multiple challenging datasets.

2. Related Work

This section introduces related work on 3D hand pose estimation, hand pose and shape recovery, and dense shape representation. Below, we compare our contribution with prior works.

2.1. 3D Hand Pose Estimation

The task of 3D hand pose estimation aims to predict the 3D position of hand joints. Recently, estimating 3D hand pose from depth image or RGB image is well-explored. For works on depth-based hand pose estimation, please refer to

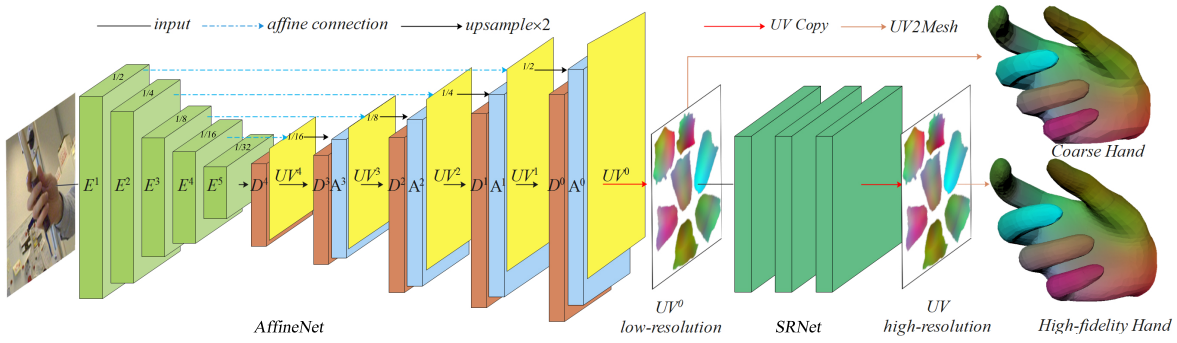

Figure 2: Overview of the propose framework. The proposed I2UV-HandNet model, which consists of the AffineNet and the SRNet, enables to predict high-precision and high-fidelity hand meshes from a single-view color image.

[2, 44]. Here, we mainly compare RGB-based hand pose estimation methods.

Because the 3D joint annotations are hard to directly obtain from the 2D images, many methods make use of the correspondence of 3D joints and their 2D projections to boost 3D pose estimation. [36] proposes to detect 2D hand keypoints of multi-view color images, and then these 2D detections are triangulated in 3D using multi-view geometry. After the emergence of many datasets with 3D annotations [46, 52], methods are explored to directly regress 3D joints from RGB images by using 3D annotations to supervise network training [52]. Then, many methods follow this strategy and improve the performance either by introducing intermediate representations [20] or using more supervision terms [7, 37]. Recently, a 3D joint is represented as three 1D heatmaps in [29], where the results are also convincing. In this paper, we only use the output 3D joints to help evaluate the performance of hand mesh modeling.

2.2. Hand Pose and Shape Estimation

Since sparse keypoints have a limited representation of 3D shapes, recent works combine sparse pose estimation with dense shape reconstruction to provide a more comprehensive shape representation. Methods in this area can be split into two categories with differences in the shape prior model used or not.

To solve the problem that 2D image lacks sufficient depth information and shape knowledge for 3D shape recovery, parametric shape models (e.g., 3DMM for face [4], SMPL for body [28], and MANO for hand [34]), which are built using 3D scan data, use low-dimension parameters to represent the complex 3D surface. Recent works [5, 15, 25, 39, 48] integrate a parametric-based hand model (MANO) with the end-to-end deep network for hand pose and shape estimation. The basic idea is to regress MANO parameters from the input image and then recover 3D joints and shape according to the regressed parameters, where 3D joints and meshes or fitted MANO parameters are used to supervise network training [5, 17, 48, 50]. There are also method which attempt to remove 3D supervision [9, 37] or recover 3D shape from 3D pose [11].

Although the parametric model brings 3D shape priors, estimating model parameters from an RGB image breaks the spatial relationship between 2D pixels. To address this issue, I2L-MeshNet [29] predicts 1D heatmaps for each mesh vertex coordinate instead of directly regressing the parameters. [14] and [27] regress per-vertex position via graph convolution networks (GCNs). Unlike them representing 3D hands in 3D space, we represent the surface of 3D hands by 2D UV maps which can be mapped from the input image in an image-to-image translation fashion like [21].

For the number of vertice of the output mesh, [27, 29] use 778 vertices and use the same mesh topology as MANO, and [14] outputs a mesh with 1,280 vertices via GCN. DeepHandMesh [30] regresses a high-fidelity mesh with 12,553 vertices, but it needs to be trained (or pre-trained) on the data captured from the controlled environment and still suffers from the highly non-linear regression problem due to it represents the hand via parameter. In this work, we propose a more general solution for high-fidelity hand reconstruction, where we input a low-resolution UV position map (MANO-level hand mesh) and output a high-resolution UV position map (high-fidelity hand mesh).

2.3. Dense Shape Representation

Although representing 3D shape via parametric models or 3D triangular mesh is straightforward and easy to employ supervision, there are works [1, 13, 43, 45, 47] which propose to represent 3D surface in a denser fashion, i.e., UV representations are introduced to represent the image-to-surface correspondences and then powerful 2D CNN can be directly utilized to learn the image-to-UV mapping. The UV representation can be divided into IUV and UV location maps, where the same position on the IUV and the RGB image shows spatial consistency, while the UV location map is inconsistent with RGB. Here, we call this inconsistency a coordinate ambiguity. The IUV representation is used in single-view 3D face reconstruction [13] and body reconstruction[42, 43]. Recently, [1, 45] combine IUV, UV position maps and SMPL model [28] to reconstruct 3D human. Even though these methods achieve good results, the

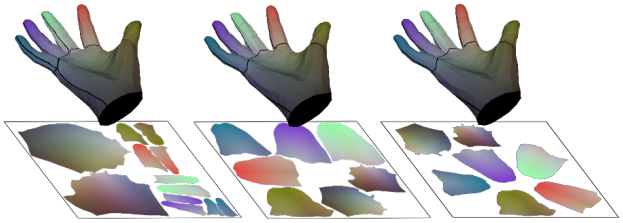

Figure 3: Three UV unfolding forms in terms of different cutting and combination strategies.

coordinate ambiguity of the RGB and UV position map is not well addressed. In this paper, we introduce UV position maps to the hand reconstruction task for the first time and propose to reduce the coordinate ambiguity via an affine connection module (Section 3.2.1).

3. Method

The proposed I2UV-HandNet model enables to represent 3D hand surface by UV position map (Section 3.1) and learn to estimate accurate and high-fidelity hand shape from a single-view image in a image-to-image translation fashion (Section 3.2 and 3.3). In the following, we describe the proposed method in detail.

3.1. 3D Hand Representation

Parametric Hand Model. MANO is a parameterized hand model learned from hand scans. It defines a mapping from pose and shape parameters to a mesh of 778 vertices and 1538 faces, where the face topology is fixed to indicate the connection of the vertices in the hand surface. From the given mesh topology, a set of 21 joints can be directly formulated from the hand mesh. Here, we use the MANO model to infer 16 joints and obtain 5 fingertips according to pre-defined vertex indexes [17].

Hand Surface as UV Position Map. Given a hand surface, such as the MANO hand mesh, we can unfold the surface into one UV map[1], which allows representing 3D surfaces as an image. Here, U and V denote the two axes of the image. The UV mapping defines the correspondence between the mesh vertices and the image pixels. Three UV mapping forms are shown in Figure 3 in terms of different cutting and combination strategies, and the ablation comparison is conducted in Section 4.6. In our pipeline, the AffineNet directly outputs a UV position map from the input image, and the SRNet outputs a UV position map from a UV position map input, and then 3D hand meshes are recovered from UV position maps via the above-defined UV mapping. To learn our model, the hand mesh annotations are transferred into UV maps to supervise the UV map prediction. Specifically, the hand mesh is spatially aligned with the corresponding RGB image using orthographic projection, so that the 3D hand mesh matches the 2D hand in the image plane. For each 3D vertex on the mesh, its 3D coordinate is mapped into the RGB channel value of a point in the UV position map [43]. Interpolation is applied to generate continuous images.

[1] https://www.autodesk.com.sg/products/maya

3.2. I2UV-HandNet

As presented in Figure 2, the proposed I2UV-HandNet model achieves accurate and high-fidelity hand shape estimation via an AffineNet to realize UV position maps prediction and an SRNet to restore high-resolution UV position maps.

3.2.1 AffineNet

To predict the UV position map of the hand shape, an encoder-decoder mechanism is adopted to map the input image into a UV image. Similar to the U-net[35], the AffineNet, consisting of a contracting path and an expansive path (as shown in Figure 2). Give a color image I with a hand in its scope, a ResNet-50 backbone [18] is used to encode the image into a series of encoded features $\{E^i|i = 1, 2, 3, 4, 5\}$ with different resolutions. In the expansive path, each step upsamples the feature map and UV map prediction along with making use of the corresponding encoded feature maps, resulting in a series of decoded features $\{D^i|i = 0, 1, 2, 3, 4\}$ and predicted UV position maps $\{I^i_{UV}|i = 0, 1, 2, 3, 4\}$:

$$\begin{cases} D^4 = f_{up}(E^5) \\ I^4_{UV} = f_{con}(D^4) \end{cases} \tag{1}$$

and

$$\begin{cases} A^3 = f_{up}(f_{ac}(\pi(I^4_{UV}), E^4)) \\ D^3 = f_{up}(D^4), \\ I^3_{UV} = f_{con}(A^3, D^3, f_{up}(I^4_{UV})), \end{cases} \tag{2}$$

and

$$\begin{cases} A^i = f_{up}(f_{ac}(\pi(I^{i+1}_{UV}), E^{i+1})), \\ D^i = f_{up}(f_{con}(D^{i+1}, A^{i+1}, I^{i+1}_{UV})), \quad i = 0, 1, 2 \\ I^i_{UV} = f_{con}(A^i, D^i, f_{up}(I^{i+1}_{UV})), \end{cases} \tag{3}$$

Here, E^i is the feature map encoded at the i-th pyramid level, A^i is the UV-aligned feature by an affine transformation, D^i is the feature map, f_{up} indicates $2\times$ up-sampling, f_{ac} indicates the affine connection operation, f_{con} indicates convolutional layers, and π indicates the projection from UV position map to image coordinate system. Smaller i indicates a higher resolution. We note that the affine connection f_{ac} aligns encoding features to decoding features through an affine-operation before connecting them, where the affine-operation, similar to the STN [22], is based on the 2D projection of each vertex coordinate in the currently predicted UV map. We provide more details in the Appendix.

3.2.2 SRNet

Since the network of our 3D hand surface is represented by UV position maps, we propose to get a more refined hand surface via a super-resolution in UV image space. We propose an SRNet to transfer the low-resolution UV position map into a high-resolution map. The network architecture of the SRNet is similar to the super-resolution convolutional neural network (SRCNN) [12], but the input and output are UV position maps instead of RGB image. After regressing a high-resolution UV position map via the SRNet, a hand mesh with higher fidelity can be reconstructed (as shown in the right part of Figure 2). More details about SRNet architecture are provided in the Appendix.

3.3. Training Objective

3.3.1 Losses of the AffineNet

To learn the AffineNet, we enforce UV alignment E_{UV}, UV gradient alignment E_{grad}, and mesh alignment E_{verts}:

$$E_{affine} = \lambda_1 E_{UV} + \lambda_2 E_{grad} + \lambda_3 E_{verts} \tag{4}$$

UV Alignment. We propose a UV alignment loss E_{UV} base on the L1 distance between the ground truth UV position map $\hat{I}_{UV}$ and the output UV position map I_{UV}:

$$E_{UV} = \left| (I_{UV} - \hat{I}_{UV}) \cdot M \right| \tag{5}$$

Here, M is UV map mask since only the valid region of the UV position map has corresponding region on the 3D hand surface.

UV Gradient Alignment. The ideal hand surface should be continuous, so does the UV position maps. To this end, we introduce a UV gradient alignment to encourage the predicted UV position map share the same gradient with the ground-truth UV position map:

$$\begin{aligned} E_{grad} = &|\partial_u (I_{UV} \cdot M) - \partial_u (I^*_{UV} \cdot M)| \\ &+ |\partial_v (I_{UV} \cdot M) - \partial_v (I^*_{UV} \cdot M)| \end{aligned} \tag{6}$$

where ∂_u and ∂_v are gradients along the U-axis and V-axis, respectively.

Mesh Alignment. Apart from E_{UV} and E_{grad} that measure the shape reconstruction in 2D UV position map space, we also introduce a mesh alignment loss E_{verts} to enforce the predicted 3D hand mesh to be closed with the ground truth one:

$$E_{verts} = \frac{1}{N_{vert}} \sum_{i=1}^{N_{vert}} |v_i - \hat{v}_i| \tag{7}$$

Here, v_i and $\hat{v}_i$ are the 3D coordinate of the i-th vertex from the output mesh and the ground truth mesh, respectively. N_{vert} indicates the number of vertices of the mesh.

Since there are multiple predicted UV position maps, we employ E_{affine} on multi-scale UV maps. When training the AffineNet, the last four UV maps are used with equal weights.

3.3.2 Losses of the SRNet

The output of the SRNet is UV position map is similar to the output of the AffineNet except that the SRNet can produce a UV map with higher resolution. Here, we adopt similar loss functions with AffineNet.

$$E_{SR} = E_{UV\_SR} + E_{verts\_SR} \tag{8}$$

Here, we replace the component in E_{UV} with the corresponding component from the SRNet to formulate E_{UV_SR}, e.g., the UV position map is replaced by the UV position map with higher resolution. Also, the E_{verts_SR} is formulated in the same manner.

4. Experiments

In this section, we first present datasets (Section 4.1) and evaluation metrics (Section 4.2), and implementation details (Section 4.3). Then, the overall performance of the proposed method and comprehensive analysis are presented (Section 4.4, 4.5 and 4.6).

4.1. Datasets

FreiHAND. The FreiHAND dataset [53] contains real-world hand data with various poses, object interactions, and varying lighting. It contains 130,240 training samples and 3,960 test samples. Each training sample contains a single-view RGB image, annotations of MANO-based 3D hand joints and mesh, as well as camera pose parameters. The result of the test set is evaluated via an online submission system[2].

HO3D. The HO3D dataset [15] is a recently released dataset that collects color images of a hand with object interactions. This dataset has 66,034 training samples, which consists of single-view RGB images, MANO-based hand joints and meshes, and camera poses. For the test set, 11,524 RGB images are provided along with the annotation of the detection bounding box. The objects in this dataset are mainly from the TCB-Video dataset [11]. The results of the test set need to be evaluated through its online submission system[3].

ObMan. The ObMan dataset [17] is a large-scale synthetic dataset containing hand-object interaction images. It contains 141,550 training samples, 6463 validation samples, and 6285 test samples. Each sample has an RGB-D

[2] https://competitions.codalab.org/competitions/21238
[3] https://competitions.codalab.org/competitions/22485

image, 3D hand joints, 3D hand mesh, object mesh as well as camera pose parameters.

YT-3D. The YouTube-3D-Hands (YT-3D) dataset collects images of various real-world hand from YouTube and annotate those images via an automated collection system [27]. The training set, which is generated from 102 selected videos, has 47,125 hand images with 3D joint and mesh annotation. The validation and test sets cover 7 videos and contain 1525 samples each.

SHS. We build the SuperHandScan (SHS) dataset using a collection of high-quality 3D hand scans via a laser scanner. The motivation of SHS is that the MANO hand mesh, which represents the hand surface via 778 vertices (with 1538 faces), can only show coarse surface information, but the UV-based method can produce a hand surface with more details. Thus, we obtain three times higher resolution hand meshes based on a collection of hand scans and the given MANO model. Specifically, we first up-sample the original MANO hand mesh from 778 vertices (with 1538 faces) to 3093 vertices (with 6152 faces) using the edge-based unpooling method as [40]. Then, the iterative closest point (ICP) algorithm is used to register the upsampled 3D mesh to the 3D point cloud (from the scanner). Our SHS dataset provides 6000 scans with dense 3D point clouds and the corresponding hand meshes. The hand mesh, which has 3093 vertices and 6152 faces, is denser than MANO hand mesh, thus can supervise SRNet to learn higher quality hand mesh.

HIC. The Hands in Action Dataset (HIC) [38] contains images of hand-object interaction. Each sample has the RGB-D image, 3D object shape, and MANO-fitted hand shape. We use all of the samples to evaluate the SRNet.

4.2. Evaluation Metrics

To evaluate the performance of the proposed method, multiple metrics are used for hand pose estimation and mesh reconstruction. The **Pose error** measures the average Euclidean distance between the predicted and the ground truth 3D joints. The **Mesh error** measures the average Euclidean distance between the predicted and the ground truth mesh vertices. The **Pose AUC** indicates the area under the curve (AUC) for the plot of the percentage of correct keypoints (PCK) and the **Mesh AUC** indicates the AUC for the plot of the percentage of correct vertices (PCV). We also compare the **F-score** [24] which is the harmonic mean of the recall and precision between two meshes given a distance threshold. We report the F-score of mesh vertices at 5mm and 15mm by F@5mm and F@15mm. Following the recent works, we compare aligned prediction results with Procrustes alignment.

We use PSNR and RMSE to evaluate the performance of hand super-resolution reconstruction. The **PSNR** indicates computes the peak signal-to-noise ratio, in decibels, between two images and is used as a quality measurement between the original and a reconstructed image. The higher the PSNR, the better the quality of the reconstructed surface. The **RMSE** (Root Mean Square Error) is the standard deviation of the residuals. In this work, we use PSNR and RMSE to evaluate the difference between the rendered depth map and the corresponding ground truth.

4.3. Implementation

We train our model on four NVIDIA Tesla V100 GPUs. Adam [23] is used to optimize the network and PyTorch [32] is used for implementation. The proposed model consists of two parts, i.e., the AffineNet and the SRNet, where the AffineNet aims to reconstruct hand mesh at MANO model level and the SRNet is designed to predict hand mesh with more detail. No image set provides both MANO-level and super MANO-level mesh annotations. Thus we adopt a stage-wise training strategy to optimize the network modules by using different data supervision for different parts, i.e., use the image dataset (such as FreiHAND) to train the AffineNet while use scan data (such as SHS) to train the SRNet.

The AffineNet is trained for 200 epochs with the batch size of 128 and the learning rate initialized to 1×10^{-4} and changed according to a Cosine Learning rate decay. The input image is cropped to $3 \times 256 \times 256$. During its training, the input image is augmented by scaling, rotation and, color channel permutation. The SRNet is trained for 100 epochs with the batch size of 512 and the learning rate is set to 1×10^{-3}. The input and output UV position map of SRNet is $3 \times 256 \times 256$. For each sample in the SHS dataset, there are two sources of SRNet's input, one is the UV of the corresponding MANO mesh, and the other is the UV map after Gaussian smoothing.

4.4. Comparison with State-of-the-art Methods

Since our SRNet part is trained by scanning data, we only compare the results of AffineNet with other methods to ensure the fairness of the comparison. We compare the proposed method with several state-of-the-art approaches [5, 11, 16, 27, 29, 53] for hand pose and shape estimation on the FreiHAND dataset. The results are shown in Table 1. In general, methods without directly regressing MANO parameters ([11, 27, 29] and our method) perform better than methods using MANO parameters regression ([5, 17, 53]). When no extra training data is used, our method surpasses all previous methods whether or not they use MANO model prior. When the additional training data (the training set of ObMan and YT-3D) is used, our method achieves better performance (see "Ours* (AffineNet)" in the table). We further plot the 3D PCK and PCV of the FreiHAND test set compared with some state-of-the-art methods [27, 29, 53], where our method shows better performance.

In the hand-object interaction scenario, we compare with

Method	Pose Error↓	Pose AUC↑	Mesh Error↓	Mesh AUC↑	F@5 mm↑	F@15 mm↑
Boukhayma *et al.* [5]	3.50	0.351	1.32	0.738	0.427	0.894
Zimmermann *et al.* [53] (Mean Shape)	1.71	0.662	1.64	0.674	0.336	0.837
Zimmermann *et al.* [53] (Mano Fit)	1.37	0.730	1.37	0.729	0.439	0.892
Hasson *et al.* [16]	1.33	0.737	1.33	0.736	0.429	0.907
Spurr *et al.* [37]	1.13	0.78	-	-	-	-
Zimmermann *et al.* [53] (MANO CNN)	1.10	0.783	1.09	0.783	0.516	0.934
Kulon *et al.* [27]	0.84	0.834	0.86	0.830	0.614	0.966
Choi *et al.* [11]	0.77	-	0.78	-	0.674	0.969
Moon *et al.* [29]	0.74	0.854	0.76	0.850	0.681	**0.973**
Ours (AffineNet)	**0.72**	**0.856**	**0.74**	**0.852**	**0.682**	**0.973**
Ours* (AffineNet)	0.68	0.865	0.69	0.862	0.706	0.977

Table 1: Comparison of main results on the FreiHAND test set. * indicates the system is trained on a combination of datasets.

Method	Pose Error↓	Pose AUC↑	Mesh Error↓	Mesh AUC↑	F@5 mm↑	F@15 mm↑
Hasson *et al.* [16]	1.14	0.773	1.14	0.773	0.428	0.932
Hasson *et al.* [17]	1.10	-	-	-	0.46	0.93
Hampali *et al.* [15]	1.07	0.788	1.06	0.790	**0.506**	0.942
Ours (AffineNet)	**0.99**	**0.804**	**1.01**	**0.799**	0.500	**0.943**
Ours† (AffineNet)	1.04	0.793	1.09	0.782	0.484	0.935
Ours∗ (AffineNet)	0.81	0.838	0.84	0.831	0.577	0.970

Table 2: Comparison of main results on the HO3D test set. † indicates cross-dataset evaluation. ∗ indicates trained use extra data.

Set	RMSE↓	PSNR↑
Input	28.12	27.58
Output (sampled)	12.06	37.74
Output	**7.68**	**39.18**

Table 3: Comparison of main results on HIC. Note that the depth maps are transferred into point clouds in world coordinates, and then the distance is computed between corresponding points.

Losses			Pose Error↓	Pose AUC↑	Mesh Error↓	Mesh AUC↑
E_{UV}	E_{grad}	E_{verts}				
✓			0.75	0.850	0.77	0.846
✓	✓		0.73	0.854	0.75	0.850
✓	✓	✓	**0.72**	**0.856**	**0.74**	**0.852**

Table 4: Ablation studies for different losses used in our method on the FreiHAND testing set.

f_{ac}	Crop Ratio	Pose Error↓	Pose AUC↑	Mesh Error↓	Mesh AUC↑
w/o	1	0.85	0.831	0.87	0.827
	3/4	0.81	0.839	0.82	0.837
	1/2	0.76	0.848	0.78	0.845
w/	1	0.77	0.847	0.79	0.843
	3/4	**0.72**	**0.856**	**0.74**	**0.852**
	1/2	0.74	0.852	0.76	0.849

Table 5: Comparison of the affine connection module f_{ac} used or not, where three crop ratios are compared in each case.

state-of-the-art methods [15, 16, 17] on the HO3D dataset in Table 2. In the condition that no extra training data is used, our method outperforms all previous methods. In addition, when the additional training data (the training set of FreiHAND, ObMan, and YT-3D) is used, our method achieves better performance (see "Ours* (AffineNet)" in the table). We also notice that our cross-dataset evaluation results ("Ours† (AffineNet)" in the table) surpass other results, where different from their training on the HO3D training set, we train the model on a combination of data from other datasets (the training data of FreiHAND, ObMan, and YT-3D). Even though most of the samples in the training set (FreiHAND, ObMan, and YT-3D) represent bare hands while the samples in the HO3D test set are hand-object interaction images, this cross-dataset evaluation still has good performance, showing the robustness and effectiveness of the proposed AffineNet.

4.5. Evaluation of Hand Super-resolution Reconstruction

The SRNet is trained on the SHS dataset, and the training detail is shown in Section 4.3. Once trained, our SRNet can be directly used for hand super-resolution reconstruction without any fine-tuning. Here, we evaluate the SRNet on FreiHAND, HO3D and HIC.

The FreiHAND and HO3D are annotated using MANO fitting, thus their ground truth meshes are as coarse as MANO hand mesh. We use the output UV position maps of AffineNet as the input of the SRNet, and the SRNet can output hand meshes with higher fidelity. The qualitative results on HO3D and FreiHAND are shown in Figure 1. For each input RGB image, we visualize the output of the AffineNet, the output of the SRNet in two viewpoints. We find that the SRNet outputs higher-resolution hand meshes than the output of the AffineNet while preserving the same pose information. The high-resolution meshes of the SRNet show smoother and more realistic skin surface.

To quantitatively evaluate the SRNet, experiments are conducted on the HIC dataset. The input of the SRNet is the UV position map converted from its MANO-fitted mesh with 778 vertices ("Input" in Table 3), and the output is the high-resolution UV position map that can be converted into a high-fidelity mesh with 3093 vertices ("Output" in the table). Besides, we also down-sample the high-fidelity mesh to 778 vertices ("Output (sampled)" in the table). In order to compare these three sets of hand surfaces, we render them as depth maps based on the viewpoint of the input image and use the depth map observation to calculate the RMSE and PSNR of the rendered depth image. Note that the background is erased by the intersection of these depth maps. As shown in Table 3, the output of SRNet ("Output" in the table) shows a better reconstruction quality to the original

Variants of our Method	AUC of PCK↑	AUC of PCV↑
UV1	0.860	0.856
UV2	0.862	0.860
UV3	**0.865**	**0.862**

Table 6: The results comparison of different UV forms in different stages.

Variants of our Method	AUC of PCK↑	AUC of PCV↑
S0	**0.865**	**0.862**
S1	0.862	0.858
S2	0.839	0.834

Table 7: The results comparison of different UV forms in different stages, S_i means the i-th pyramid level output I^i_{UV}.

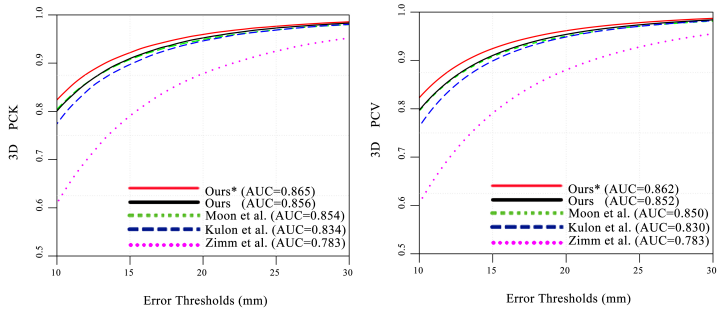

Figure 4: Comparison of 3D PCK and 3D PCV on the FreiHAND dataset. The proposed method is superior to [27], [29] and [53].

depth map. We also notice that the down-sampled output ("Output (sampled)" in the table) obtains higher accuracy than the mesh fitted by MANO ("Input" in the table), even though they have the same mesh resolution.

4.6. Ablation Study

Effect of Each Loss Function. As presented in Table 4, we give evaluation results of AffineNet on the FreiHAND test set of settings with different losses used during the training. From the table, we can see that the best performance is achieved when all proposed loss functions are used. More results are shown in the Appendix.

Effect of the Affine Connection. As presented in Table 5, we give evaluation results on FreiHAND of settings with the affine connection module f_{ac} used or not, where three crop ratios are compared in each case. Here, the cropping operation is designed to compare the effect of different foreground and background ratios on the reconstruction result. For example, when the crop ratio is 1/2, the image is center cropped by 1/2 of the width and height and then resized into the original size. In Table 5, for each crop ratio, we find the "w/ f_{ac}" gets better performance than "w/o f_{ac}" and it gets the best performance when the crop ratio is 3/4. Therefore, we choose 3/4 as the crop ratio in other experiments. In this case, f_{ac} results in an 11.1% reduction in pose error and a 9.8% reduction in mesh error.

Effect of UV unfolding forms. The UV map is obtained by unfolding the hand mesh. Thus, the valid part of the UV map is dependent on the crop trajectory of the surface (i.e., the black lines on the 3D surface in Figure 3). We compare three different UV position maps using different crop and combine schemes, and indicate UV1, UV2 and, UV2 from left to right in Figure 3. UV1 separates the front and back of the hand and the area on the UV position map of each piece is proportional to the area of the mesh surface. UV2 and UV3 don't separate each finger and the area on the UV position map of each piece is proportional to the vertex number of the mesh surface, where each piece has a different position in UV space. In Table 6, we give evaluation results on the FreiHAND test set while using a combined training set (refer to the train data of FreiHAND, ObMan, and YT-3D). Despite different UV position maps are used, the performance of hand pose and shape estimation is similar, where less than 0.7% difference on AUC of PCK/PCV is shown. The results show that our method is robust to the UV position map designing. In this paper, we use UV3 as the template UV position map.

Comparison of results from UV maps of different stages. As illustrated in Section 3.3.1, the supervision is used on multiple-scale UV maps. Thus, the mesh can be recovered from each UV map. Here, we compare the mesh prediction results of the last three UV maps (S0, S1, and S2) on the FreiHAND test set while using a combined training set. S0 indicates the full resolution UV map prediction, while S1 refers to the second last output with 1/2 of the full resolution and S2 refers to the third last output with 1/4 of the full resolution. As shown in Table 7, from low-resolution S2 to high-resolution S0, the AUC of PCK and PCV show obvious improvement.

5. Conclusion

We have presented a novel I2UV-HandNet approach for accurate and high-fidelity 3D hand reconstruction from a single color image. The proposed UV position map enables representing a 3D hand surface in an image style. For accurate hand pose and shape estimation from the monocular images, we present an AffineNet to predict the UV position map from the RGB input. In AffineNet, a hierarchical coarse-to-fine regressing architecture is designed with a novel affine connection module that can resolve the coordinates ambiguity between the RGB image and the UV map. The proposed AffineNet achieves state-of-the-art performances on multiple challenging datasets. For high-fidelity hand shape reconstruction, we present an SRNet to restore a high-resolution UV position map from a low-resolution one. The proposed SRNet is not likely to be affected by the reconstruction method, and can robustly restore a high-fidelity hand from the inputted coarse shape. As for the future study, the UV-based hand representation can be extended to more complex joint hand-hand or hand-object reconstruction tasks, or the architecture can be modified for enabling sparse/weak supervision.

References

[1] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *International Conference on Computer Vision*, 2019.

[2] Anil Armagan, Guillermo Garcia-Hernando, Seungryul Baek, Shreyas Hampali, Mahdi Rad, Zhaohui Zhang, Shipeng Xie, MingXiu Chen, Boshen Zhang, Fu Xiong, et al. Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3d hand pose estimation under hand-object interaction. In *European Conference on Computer Vision*, 2020.

[3] Vassilis Athitsos and Stan Sclaroff. Estimating 3d hand pose from a cluttered image. In *Conference on Computer Vision and Pattern Recognition*, 2003.

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Annual Conference on Computer Graphics and Interactive Techniques*, 1999.

[5] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[6] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[7] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *European Conference on Computer Vision*, 2018.

[8] 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*, 2019.

[9] 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 *Conference on Computer Vision and Pattern Recognition*, 2021.

[10] Yujin Chen, Zhigang Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and Junsong Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *IEEE Transactions on Image Processing*, 30:4008–4021, 2021.

[11] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, 2020.

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[13] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision*, 2018.

[14] 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 *Conference on Computer Vision and Pattern Recognition*, 2019.

[15] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[16] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[17] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[19] Markus Höll, Markus Oberweger, Clemens Arth, and Vincent Lepetit. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces*, 2018.

[20] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *European Conference on Computer Vision*, 2018.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[22] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 2017–2025, 2015.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2014.

[24] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 2017.

[25] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision*, 2019.

[26] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[27] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 2015.

[29] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, 2020.

[30] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *European Conference on Computer Vision*, 2020.

[31] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (TOG)*, 38(4):1–13, 2019.

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems*, 2019.

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[34] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 2017.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[36] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[37] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European Conference on Computer Vision*, 2020.

[38] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193, 2016.

[39] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.

[40] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *European Conference on Computer Vision*, 2018.

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[42] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *International Conference on Computer Vision*, 2019.

[43] Pengfei Yao, Zheng Fang, Fan Wu, Yao Feng, and Jiwei Li. Densebody: Directly regressing dense 3d human pose and shape from a single color image. *arXiv preprint arXiv:1903.10153*, 2019.

[44] 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 *Conference on Computer Vision and Pattern Recognition*, 2018.

[45] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[46] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016.

[47] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[48] 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*, 2019.

[49] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *International Conference on Computer Vision*, 2017.

[50] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[51] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, 2017.

[52] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *International Conference on Computer Vision*, 2017.

[53] 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 *International Conference on Computer Vision*, 2019.