Consistent-Resolution Network for 3D Hand Shape Estimation from a Single RGB Image

Qi Wu¹², Joya Chen¹, Zhiming Yao², Xu Zhou², Jianguo Wang²³, Shaonian Wang²³ and Xianjun Yang²

¹Institution of Physical Science and Information Technology, Anhui University, Hefei City, Anhui Province, China
²Institution of Intelligent Machines, Chinese Academy of Sciences, Hefei City, Anhui Province, China
³Department of Computer Science and Engineering, University of Science and Technology of China, Hefei City, Anhui Province, China
Email: xjyang@iim.ac.cn

Abstract. We propose a novel method for 3D hand shape estimation from a single RGB image. Most exiting methods leverage a deep network to extract a low-resolution representation to estimate 3D coordinates, which always leads to the loss of spatial information. In contrast, we present a Consistent-Resolution Network (CRNet) to extract the same resolution representation as the original image, thus preserve more details about spatial information. Specifically, we introduce the recent high-resolution network (HRNet) to generate high-resolution feature maps, which can attain high-resolution representation of the original image. Then, we design a deconvolution module to recover this map to the size of the original image. Therefore, we can directly leverage this feature to learn the precise 2D shape and the depth map, and transfer them into 3D coordinates in the camera space. Through extensive experiments on a large real-world dataset FreiHAND, we show that our proposed method can predict precise and suitable 3D hand shape from a monocular view.

1. Introduction
3D human hand analysis, which is an extremely important research field in graphics communities and computer vision, has abundant applications in various fields, including VR/AR, human-machine interaction, and robotics [1, 2]. With the emergence of vast economical depth cameras, previous works mainly concentrated on estimating 3D hand pose from depth images [3-5]. Since depth cameras do not work under bright light and RGB cameras are more common, some latest researches start estimating sparse 3D hand joint locations from single RGB images but ignore dense 3D hand shape [6-10]. Compared with existing studies on 3D hand pose estimation, 3D hand shape is a more ample and comprehensive representation for image understanding.

However, due to the inherent depth ambiguity of the monocular setting, self-occlusions, varying hand shapes, and self-similarity, 3D hand shape estimation is very challenging, especially from a single RGB image. Various deep learning methods have approached these problems, but the performance of their results depends on the amount of the training data. Since the ground-truth annotation for 3D hand shape in real-world RGB images is extremely laborious and cumbersome, some works [6, 11] have adopted synthetic datasets for training. Nevertheless, models trained on synthetic images generalize poorly to real images on account of the domain gap [8], which causes...
distorted hand shape. Meanwhile, current deep learning methods [12, 13] utilize a deep network to extract a low-resolution representation to estimate 3D coordinates, which always leads to the loss of spatial information.

In this work, we propose a Consistent-Resolution Network (CRNet) to extract the same resolution representation as the original image, thus preserve more details about spatial information. Specifically, we introduce the recent high-resolution network (HRNet [14]) to generate high-resolution feature maps, which can attain high-resolution representation of the original image. Then, we design a deconvolution module to recover this map to the size of the original image. Therefore, we can directly leverage this feature to learn the precise 2D shape and the depth map, and transfer them into 3D coordinates in the camera space.

To validate the effectiveness of our proposed method, we conduct extensive experiments on a large real-world dataset FreiHAND [15], which consists of real images, various hand pose and shape, and hand-object interactions. Experimental results show that our proposed method can predict precise and suitable 3D hand shape from monocular.

2. Related Work
Now, we present the review for previous works on 3D hand shape estimation, consisting of based depth images and based RGB images.

2.1. 3D Hand Shape Estimation from Depth Images
Many works [12, 16-20] propose to estimate 3D hand shape form depth images due to the widespread commodity depth cameras. The previous methods [16-19] mainly fit a deformable hand model onto depth images with an iterative optimization. With the progress in deep learning, Malik et al. [12] proposed a first deep neural network for hand pose and shape estimation from a single depth image, which is a 2D CNN-based approach that estimates shapes directly from 2D depth maps. Recently, Malik et al. [20] proposed another structured weakly-supervised deep learning-based approach using a single depth image. However, the performance of these methods is restricted by the inherent drawbacks of depth sensors, which do not work under bright sunlight, have a high power consumption and people have to close to the sensor.

2.2. 3D Hand Shape Estimation from a Single RGB Image
Recently, researchers started to focus on 3D hand shape estimation from a single RGB image. Panteleris et al. [21] estimated 2D joint locations by fitting a 3D hand model which is controlled by 27 hand pose parameters. Therefore it is hard to be suitable for various hand shapes. Ref. [22] fitted a generic hand model to the predictions via an iterative optimization based approach, which is not time-efficient and requires hand-crafted energy functionals. Refs. [23, 24] proposed to regress the parameters of a deformable hand mesh model from the input image in an end-to-end manner. Ge et al. [13] directly regressed a hand shape using a GraphCNN, but a special dataset with ground truth and meshes is required, which is hard to construct. Although these works achieve appealing results, they almost extract a low-resolution representation to estimate 3D coordinates, which always leads to the loss of spatial information.

3. Methodology
As shown in figure 1, given an image, our CRNet can output the same resolution representation (i.e. feature map) as the original image, and predict accurate 3D vertexes based on the feature map. We introduce the well-known HRNet [12] as the backbone to obtain high-resolution representation, and design downsampling and upsampling modules to improve the computational efficiency. The learning for 3D hand shape is divided into two steps: (1) learning 2D vertexes and corresponding depth; (2) obtaining 3D vertexes in the camera space. In the following, we will introduce the pipeline of our CRNet in details, and show how it learns 3D hand shape from a single RGB image.
3.1. Consistent-Resolution Network (CRNet)

HRNet Backbone. HRNet (High-Resolution Network) [14] maintain high-resolution representation through the network forwarding process, rather than introducing an encoder-decoder framework. We believe that this representation is significant for accurate 3D hand shape estimation. Specifically, 3D hand shape estimation relies on 3D coordinates prediction of the hand, which requires accurate spatial information. The encoder-decoder framework will lose spatial information in the high-to-low resolution branch, but may not recover the information in the low-to-high resolution branch.

However, when the size of the image is too large, maintaining the same resolution representation as the original image will be time-consuming. Therefore, we use HRNet as the backbone network in our approach, but design a scaling module with HRNet to improve computational efficiency.

Scaling Module. We design this module to control the size of the feature map in HRNet backbone. Firstly, the input image will be passed to $\alpha$ convolutional layers with stride 2. Then, the downsampling rate of the feature map will be $2^\alpha$. Similarly, the feature map after HRNet will be passed to $\alpha$ deconvolutional layers with stride 2. Therefore, the size of the obtained feature map will be the same as the original image. Our proposed scaling module can ensure the HRNet backbone to do efficient computation. Note that the mentioned convolutional layer is a simple 3×3 kernel with stride 2, using the ReLU activation, of which is the transpose version of the deconvolutional layer.

3.2. CRNet for 3D Hand Shape Learning

Training. As shown in figure 1, the obtained consistent-resolution feature map will be passed to two 3x3 convolutional layers to predict 2D vertexes and depths, respectively. Note that the 2D vertexes and depths annotation can be computed from 3D vertexes annotation in camera space. Specifically, given the camera matrix

$$K = \begin{bmatrix} f_x & 0 & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

and 3D vertexes $(x, y, z)$, we can compute 2D vertexes $(u, v)$ and the depth $d$:

$$u = f_x x + x_0 z \quad (2)$$

$$v = f_y y + y_0 z \quad (3)$$

$$d = z \quad (4)$$
They can be directly learned in our network. We use binary cross-entropy loss with sigmoid activation to learn 2D vertexes, and use smooth L1 loss to learn depth. The whole loss function of the network is:

$$L = \sum_{i=0}^{WH} 1_{y_i=1} \log(\sigma(x_i)) + 1_{y_i=0} \log(1 - \sigma(x_i)) + 1_{y_i=1} L_{smooth}(d_i, \hat{d}_i)$$

where $W$ and $H$ is the width and height of the feature map, respectively. $y_i$ is the label of i-th point in the feature map, where the ground-truth vertexes corresponds $y_i = 1$, otherwise $y_i = 0$. $\sigma(x_i)$ is the score with sigmoid activation of i-th point in the 2D vertexes prediction map. We can see that the learning of 2D vertexes uses the cross-entropy loss to train. $L_{smooth}$ denotes the smooth L1 loss, and $d_i, \hat{d}_i$ is the depth prediction, depth annotation, respectively. Once the 2D vertexes and the depth prediction maps are obtained, we can transfer them to 3D vertexes in the camera space, as we illustrate in follows.

**Inference.** During inference, we use CRNet to obtain feature maps of 2D vertexes and depths, and transfer them to 3D vertexes in the camera space. This is an inverse process of the annotation production during training. Specifically, given the 2D vertexes $(u, v)$ and the depth $d$, we can compute 3D vertexes $(x, y, z)$:

$$x = (u - x_0 d) / f_x$$  \hspace{1cm} (6) \\
y = (v - y_0 d) / f_y$$  \hspace{1cm} (7) \\
z = d$$  \hspace{1cm} (8)

**4. Experiments**

**4.1. Implementation Details**

**Dataset.** We evaluate our proposed method on the public RGB dataset FreiHAND [15], which was created with the multi-view setup, consisting of real images, various hand pose and shape, and hand-object interactions. FreiHAND is a large-scale real-world dataset containing 32560 samples with the ground truth of 778 hand shape vertexes. We randomly select 5000 samples as the evaluation set and report mesh error. The training samples were recorded with a green screen background allowing for background removal.

**Hyper-parameters.** We train our network in 60 epoches by Adam, with batch size 32 on 4 NVIDIA TITAN GPUs. Our CRNet needs to tune the hyper-parameter $\alpha$, which means the scaling factor of the original image (See Section 3.1). The inference speed is measured on a single TITAN X GPU.

**4.2. Results**

As shown in table 1, when the scaling module is not used ($\alpha = 0$), the network presents the slowest inference speed. By adding the convolutional and deconvolutional layers, the inference will be faster, with little loss in mesh error. But when the resolution of the feature map is too small ($\alpha = 3$), the mesh error will be obviously improved. Therefore, we recommend $\alpha = 2$ in training CRNet.

**Table 1.** The performance of CRNet in different settings.

| Metric | $\alpha = 0$ | $\alpha = 1$ | $\alpha = 2$ | $\alpha = 3$ |
|--------|-------------|-------------|-------------|-------------|
| Mesh error | 0.89 | 0.89 | 0.90 | 1.23 |
| Speed (ms) | 38.23 | 21.96 | 18.55 | 18.26 |
4.3. Visualization  
We show some quantitative results of 3D hand shape estimation with the ground-truth MANO shape in figure 2. As we can see, the proposed CRNet is able to predict accurate and reasonable 3D hand shape.

![Figure 2](image)  
*Figure 2. The visualization results of our proposed CRNet.*

5. Conclusion  
In this paper, we proposed the novel Consistent-Resolution Network (CRNet) for estimating 3D hand shape from a single RGB image. Our method adopts the High-Resolution Network (HRNet) as a backbone to maintain high-resolution representation. To improve computational efficiency, we design a scaling model, which inserts some convolutional layers before HRNet and inserts corresponding deconvolutional layers after HRNet. Hence, our CRNet extracts the same resolution representation as the original image, so that we can learn accurate 2D vertexes and depths of the hand, and convert them to 3D vertexes of the hand in the camera space. Experiments on FreiHAND validated the effectiveness of our proposed method.

Acknowledgments  
This study was supported by National Natural Science Foundation of China (NSFC) for Youth (grant number 61701483) and Research & Development Project of Scientific Research Equipment of Chinese Academy of Sciences (grant number YJKYYQ20170011).

References  
[1] Jang Y, Noh S T, Chang H J, et al. 2015 3D finger cape: Clicking action and position estimation under self-occlusions in egocentric viewpoint *IEEE Transactions on Visualization and Computer Graphics* 21 (4) 501-510.

[2] Hürst W and Van Wezel C 2013 Gesture-based interaction via finger tracking for mobile augmented reality *Multimedia Tools and Applications* 62 (1) 233-258.

[3] Ge L, Liang H, Yuan J, et al. 2017 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 1991-2000.

[4] Wan C, Probst T, Van Gool L, et al. 2018 Dense 3D regression for hand pose estimation *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 5147-5156.

[5] Poier G, Schinagl D and Bischof H 2018 Learning pose specific representations by predicting different views *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 60-69.

[6] Zimmermann C and Brox T 2017 Learning to estimate 3d hand pose from single RGB images *Proceedings of the IEEE International Conference on Computer Vision* pp 4903-4911.
[7] Spurr A, Song J, Park S, et al. 2018 Cross-modal deep variational hand pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 89-98.

[8] Mueller F, Bernard F, Sotnychenko O, et al. 2018 Generated hands for real-time 3D hand tracking from monocular RGB. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 49-59.

[9] Cai Y, Ge L, Cai J, et al. 2018 Weakly-supervised 3D hand pose estimation from monocular RGB images. Proceedings of the European Conference on Computer Vision pp 666-682.

[10] Iqbal U, Molchanov P, Breuel Juergen Gall T, et al. 2018 Hand pose estimation via latent 2.5D heatmap regression. Proceedings of the European Conference on Computer Vision pp 118-134.

[11] Simon T, Joo H, Matthews I, et al. 2017 Hand keypoint detection in single images using multiview bootstrapping. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition pp 1145-1153.

[12] Malik J, Elhayek A, Nunnari F, et al. 2018 DeepHps: End-to-end estimation of 3D hand pose and shape by learning from synthetic depth. 2018 International Conference on 3D Vision (3DV) (IEEE) pp 110-119.

[13] Ge L, Ren Z, Li Y, et al. 2019 3D hand shape and pose estimation from a single RGB image. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 10833-10842.

[14] Sun K, Xiao B, Liu D, et al. 2019 Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 5693-5703.

[15] Zimmermann C, Ceylan D, Yang J, et al. 2019 FreiHAND: A dataset for Markerless capture of hand pose and shape from single RGB images. Proceedings of the IEEE International Conference on Computer Vision pp 813-822.

[16] Khamis S, Taylor J, Shotton J, et al. 2015 Learning an efficient model of hand shape variation from depth images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 2540-2548.

[17] Joseph Tan D, Cashman T, Taylor J, et al. 2016 Fits like a glove: Rapid and reliable hand shape personalization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 5610-5619.

[18] Tkach A, Tagliasacchi A, Remelli E, et al. 2017 Online generative model personalization for hand tracking. ACM Transactions on Graphics (ToG) 36 (6) 1-11.

[19] Remelli E, Tkach A, Tagliasacchi A, et al. 2017 Low-dimensionality calibration through local anisotropic scaling for robust hand model personalization. Proceedings of the IEEE International Conference on Computer Vision pp 2535-2543.

[20] Malik J, Elhayek A and Stricker D 2019 WHSP-Net: A weakly-supervised approach for 3D hand shape and pose recovery from a single depth image. Sensors, 19 (17) 3784.

[21] Panteleris P, Oikonomidis I and Argyros A 2018 Using a single RGB frame for real time 3D hand pose estimation in the wild. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (IEEE) pp 436-445.

[22] Xiang D, Joo H and Sheikh Y 2019 Monocular total capture: Posing face, body, and hands in the wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 10965-10974.

[23] Zhang X, Li Q, Mo H, et al. 2019 End-to-end hand mesh recovery from a monocular RGB image. Proceedings of the IEEE International Conference on Computer Vision pp 2354-2364.

[24] Baek S, Kim K I and Kim T K 2019 Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 1067-1076.