Gesture Pose Estimation Based on Pseudo-3D Convolutional Neural Networks (Postprint)

Authors: Zhang Hongyuan, Yuan Jiazheng, Liu Hongzhe, Yuan Chunfeng, Wang Xueqiao, Deng Zhifang

Date: 2019-01-28T14:33:27+00:00

Abstract

Most existing deep learning-based hand gesture pose estimation methods employ standard three-dimensional convolutional neural networks to extract three-dimensional features and estimate hand joint coordinates. The features extracted by this approach lack multi-scale information of the hand, limiting the accuracy of hand gesture pose estimation. Additionally, due to the substantial computational cost and memory requirements of three-dimensional convolutional neural networks, these methods often struggle to meet real-time requirements. To overcome these limitations, we propose simulating three-dimensional convolution through a cascade of spatial filters and depth filters, thereby reducing the number of network parameters. Simultaneously, hand gesture pose features are extracted at multiple scales and integrated to fully utilize the three-dimensional information of gestures. Experiments demonstrate that the proposed method can effectively improve the accuracy of hand gesture pose estimation, reduce model size, and run at speeds exceeding 119 fps on a computer equipped with a single GPU.

Full Text

Preamble

Hand Pose Estimation Using Pseudo-3D Convolutional Neural Network

Zhang Hongyuan[1], Yuan Jiazheng[2]†, Liu Hongzhe[1], Yuan Chunfeng[3], Wang Xueqiao[1], Deng Zhifang[1]
(1. Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China;
2. Beijing Open University, Beijing 100081, China;

3. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Most existing deep learning-based hand pose estimation methods employ standard three-dimensional convolutional neural networks (3D-CNN) to extract 3D features and estimate hand joint coordinates. However, the features extracted by these methods lack multi-scale information of the hand, which limits estimation accuracy. Moreover, due to the enormous computational cost and memory requirements of 3D CNNs, these methods often struggle to meet real-time demands. To overcome these limitations, we propose simulating 3D convolutions through cascaded spatial and depth filters, thereby reducing network parameters. Simultaneously, we extract and integrate gesture pose features across multiple scales to fully utilize 3D information. Experiments demonstrate that our method effectively improves hand pose estimation accuracy, reduces model size, and runs at over 119 fps on a computer with a single GPU.

Keywords: hand pose estimation; pseudo-3D convolutional neural network; 3D features; depth image; deep learning

0 Introduction

Vision-based hand pose estimation research has made significant progress in recent years. As a core technology for human-computer interaction, it provides users with a natural interaction modality. Since depth images can effectively address issues such as complex background interference present in monocular RGB input, hand pose estimation tasks have almost exclusively shifted to using only depth data as input [1~6]. Furthermore, deep learning has transformed approaches to vision problems, and the use of deep neural networks has become standard practice in hand pose estimation methods [7~9].

Among the numerous pose estimation methods based on deep neural networks, depth maps are often treated as 2D images and fed into 2D convolutional neural networks (CNNs) to output 3D joint positions [9,10], hand model parameters [11], or heatmaps [12]. Intuitively, due to the lack of 3D spatial information, image-based features extracted by 2D CNNs are not well-suited for 3D hand pose estimation. In light of this, several 3D CNN-based methods have been proposed recently [1,13,14]. However, these methods simply apply 3D CNNs for feature extraction without fully leveraging 3D information. Additionally, training 3D networks requires substantial computational cost, and model size nearly doubles compared to 2D CNNs. To achieve real-time performance, only shallow network structures can be used, which significantly compromises pose estimation effectiveness.

Recently, to address the massive computational cost and memory requirements of 3D CNNs, reference [15] proposed a novel network architecture called Pseudo-3D Residual Networks (P3D ResNet). This innovative module design substantially reduces model size while maintaining accuracy. Reference [16] introduced a new “stacked hourglass” network for human pose estimation that extracts and

merges human pose features at different scales, thereby significantly improving pose estimation accuracy. Inspired by these works, we propose a hand pose estimation method based on pseudo-3D convolutional neural networks. The overall network architecture is shown in [Figure 1: see original paper]. First, the depth map of hand gestures is encoded into a 3D volumetric representation, and the hand region is segmented from this representation and fed into a complete network composed of basic pseudo-3D residual modules, which finally outputs the 3D coordinates of hand joints. The advantages of our method can be summarized as follows:

a) Using an improved volumetric representation method for hand pose, we train a simple CNN to obtain more accurate hand regions, eliminating the influence of invalid regions.

b) Using pseudo-3D convolutions to replace standard 3D convolutions, we significantly reduce model size and accelerate hand pose estimation speed.

c) Using a 3D "hourglass" structure network, we extract and fuse multi-scale features of hand pose, fully utilize 3D information, and improve hand pose estimation accuracy.

1.1 Hand Pose Estimation from Depth Images

Methods for hand pose estimation from depth images can be categorized into model-based methods, data-driven methods, and hybrid approaches. Model-based methods typically predefine a hand model and minimize a loss function to match the hand model with the input depth image. Common optimization methods include Iterative Closest Point (ICP) [17], Particle Swarm Optimization (PSO) [18], or their combination [4]. Since these methods usually require temporal information, they are more dependent on hand model initialization and errors tend to accumulate during pose estimation.

Data-driven methods directly localize hand joints from input depth maps. Inspired by methods in the human pose estimation domain [19], references [20,21] used random forest-based methods and their improvements as discriminative models, achieving accurate and fast performance. However, limited by hand-crafted features, random forest-based methods currently struggle to surpass CNN-based pose estimation methods. Our work is related to CNN-based data-driven methods. Reference [7] first proposed estimating 2D heatmaps for each hand joint via CNN to localize hand joints. References [22,23] proposed a 2D Region Ensemble Network (REN) to accurately estimate 3D joint coordinates.

Reference [16] proposed a novel "Stacked Hourglass Networks" (SHN) that extracts and integrates image features at various scales to accurately estimate 2D joint coordinates, achieving significant success in human pose estimation. Reference [1] innovatively introduced 3D CNNs into hand pose estimation tasks, utilizing 3D information from depth maps to directly estimate 3D coordinates

of hand joints. Reference [13] employed a multi-task cascaded network combining 2D joint detection with 3D pose estimation, simultaneously leveraging 2D and 3D information from depth maps for hand pose estimation. These methods all use simple 3D CNNs to extract hand pose features but do not fully utilize information across different scales in depth maps. Our method leverages the advantages of both SHN and 3D CNN, extracting and integrating 3D features from multiple scales for hand pose estimation.

Hybrid methods combine model-based and data-driven approaches. Reference [24] trained a feedback loop composed of multi-level networks, including a discriminative network for initial pose estimation, a generative network for pose synthesis, and a pose update network that improves pose estimation through multiple iterations. Reference [25] used two deep generative models with a shared latent space and trained a discriminator to estimate occluded parts of hand gestures. Our work focuses on single-stage, one-shot complete hand pose estimation to more effectively exploit the potential correlations between hand joints.

1.2 Pseudo-3D Convolutional Neural Networks

3D CNNs have been successfully applied to extract 3D features from data such as depth maps and CAD models for tasks including 3D scene reconstruction [26], 3D object detection [27], and object recognition [28]. However, shallow 3D CNNs struggle to obtain effective features, while training deep 3D CNNs requires high computational cost and memory demand.

To address these issues, the P3D ResNet proposed in reference [15] uses a combination of spatial and temporal convolution filters to simulate spatio-temporal convolution filters. This combination can be considered a pseudo-3D CNN, which significantly reduces model parameters and shortens training time while improving video analysis performance. This work has been successfully applied to various video-related tasks, but they did not focus on 3D hand pose estimation. Our work uses spatial and depth convolution filters to simulate 3D convolutions, extracting 3D features of hand pose while reducing neural network parameters to meet real-time requirements in practical applications. Implementation details will be described in Section 3.2.

2 Method Overview

To fully utilize 3D information across various scales in hand pose depth maps and accelerate hand pose estimation speed, we propose a novel pseudo-3D stacked hourglass network. The hourglass network facilitates multi-scale feature extraction, while the pseudo-3D structural design effectively reduces computational cost and spatial requirements for network training. First, a single depth image containing hand pose is converted into voxel form through volumetric representation, and the hand region is segmented and fed into the network. Then, through multiple 3D convolution, 3D pooling, and 3D deconvolution operations,

the network extracts multi-scale 3D features from the volumetric representation and finally regresses the 3D coordinates of hand joints. To make the network model more robust to different hand sizes and camera viewpoints, our method performs data augmentation on the voxelized hand pose. The overall network structure is shown in [Figure 1: see original paper], where the numbers below each module indicate the "size@channels" of the input (top) and output (bottom) feature maps, with N^3 representing N × N × N.

2.1 Volumetric Representation of Hand Pose

Volumetric Representation. The goal of encoding hand pose into a volumetric representation is to represent the 3D volume of hand pose in space as comprehensively as possible from depth images. Our work improves upon the occupancy grid model proposed in reference [28] by adopting a new hand region acquisition method. First, each pixel in the depth map is reprojected into 3D space based on its depth value, and then this space is divided into voxel grids according to a predefined voxel resolution. As shown in Equation (1), if a voxel grid contains depth points, the voxel value is set to 1; otherwise, it is set to 0.

To determine the hand region reference point, we employ a new shallow network based on a 2D ResNet [29] to learn the reference point offset vector for the hand's middle finger metacarpophalangeal (MCP) joint. Compared to the shallow CNN used in reference [24], our method can extract more effective features from hand depth maps and regress a precise reference point for the hand region. A bounding box is then drawn centered at this point to remove the influence of invalid regions.

Data Augmentation. In practical applications, hand gestures exhibit significant variations in hand shape size and viewing angle. To make the model more robust, we perform data augmentation on the training data. Specifically, the segmented hand region is randomly rotated in the XY space and randomly scaled and translated in 3D space, where the rotation angle ranges in [-45°, 45°], the scaling factor ranges in [0.7, 1.3], and the translation pixels range in [-10, 10]. Both the original dataset and the augmented dataset are used for training.

2.2.1 Basic Modules

The pseudo-3D hourglass structure design we propose can extract features at different scales and then fuse them. Local features at small scales are crucial for estimating hand joint positions, while global features at large scales can fully exploit the potential overall correlations between hand joints, thereby improving hand pose estimation accuracy. To this end, we primarily use the following three basic modules to build our network model.

3D Convolution Module. This module consists of a standard 3D convolutional layer, 3D batch normalization, and activation function (ReLU), mainly applied at both ends of the network for shallow feature extraction from input

and filtering of integrated multi-scale features.

3D Deconvolution Module. This module comprises a 3D deconvolutional layer, 3D batch normalization, and activation function (ReLU), located in the latter half of the hourglass structure and primarily used for upsampling feature maps.

Pseudo-3D Residual Module. ResNet [29] consists of many residual blocks, where each residual block can be represented as $x_{t+1} = x_t + R(x_t)$, where x_t and x_{t+1} represent the input and output of the t-th residual unit, and R is a nonlinear residual function. The main idea of ResNet is to fit the residual function rather than directly learning the nonlinear mapping function.

To reduce neural network parameters and extract more effective features, as shown in [Figure 2: see original paper], we adopt a new filter combination that decomposes the original 3D convolutional filter $(3\$\times 3\times 3 convolution) in the 3D residual module([Figure 2: see original paper](a)) into a spatial convolutional filte$ convolution), as shown in Figure 2: see original paper, where S extracts spatial features of hand pose and D further extracts depth features of hand pose. The two convolutional filters are cascaded to form the basic pseudo-3D residual module, which can be represented as $x_{t+1} = x_t + D(S(x_t))$, where x_t and x_{t+1} represent the input and output of the t-th residual unit, and S and D are two nonlinear residual functions on the same path.

To simplify the model learning process and accelerate training speed, we add batch normalization layers and activation functions to all basic modules, set the convolution kernel size in the 3D convolution module to $3\$\times 3\times 3, and set the kernel size in the 3D pooling layer and 3D deconvolution module to 2\times 2\times\$2$ with a stride of 2.

2.2.2 Pseudo-3D Hourglass Structure

To address the problem that existing 3D structure networks cannot fully utilize multi-scale features, we propose a novel pseudo-3D hourglass structure that extracts features from various scales and fuses them to improve pose estimation accuracy. The volumetric representation of hand pose undergoes multiple consecutive pooling operations to obtain feature maps at different scales, and pseudo-3D residual modules extract features f_k from these multi-scale feature maps. To integrate features from different scales, the smallest-scale feature map undergoes multiple 3D deconvolution modules for upsampling and fuses the extracted features f_k from the feature extraction stage to assist upsampling. Finally, features from all scales are integrated to fully utilize 3D information of hand pose and determine 3D coordinates of hand joints. The pseudo-3D hourglass structure is shown in [Figure 3: see original paper], where numbers below each module indicate the "size@channels" of input (top) and output (bottom) feature maps, with N^3 representing N × N × N.

In the hourglass structure, 3D pooling layers reduce the spatial dimensions of

feature maps, while pseudo-3D residual modules increase the number of channels and extract 3D features of hand pose. These features are passed along two paths: one undergoes multiple pooling and pseudo-3D convolution operations, while the branch path undergoes a simple filtering operation once and then fuses with the upsampled small-scale features. When feature maps reach the lowest resolution, after extracting 3D features through 3 consecutive pseudo-3D residual modules, we use 3D deconvolution modules for upsampling and fuse them with features from the branch path. Since the hourglass structure is symmetric, corresponding branch path features assist upsampling during each feature fusion, ensuring that the overall input and output of the hourglass structure have the same dimensions.

Regarding the overall network structure, as shown in [Figure 1: see original paper], the volumetric representation of hand pose is filtered by a 3D convolution module with a $7\$\times7\times7 kernel, then downsampled by a 3D pooling layer to a size suitable for input to the pseudo-3D hourglass structure. After passing through 3 consecutive pseudo-3D residual modules, it enters the hourglass$ 3D convolutions are applied to process the integrated features, followed by 2 fully connected layers to regress the 3D coordinates of hand joints.

2.3 Network Training

Due to the pseudo-3D structure design, our method substantially reduces the computational cost and spatial requirements for network training. During training, the network does not load pretrained models, and the loss function L is computed using mean squared error as shown in Equation (4):

$$L = \frac{1}{M}\sum_{m=1}^{M}\|G_m - C_m\|^2$$

where G_m and C_m are the ground-truth and estimated 3D coordinates of the m-th hand joint, respectively, and M represents the number of joints per hand.

The network is trained and tested using the Torch7 framework on a single NVIDIA Titan X GPU. All weight parameters in the network are initialized using a zero-mean Gaussian distribution and updated by the RMSProp optimizer with a learning rate of 10^{-3} and a mini-batch size of 8. According to the actual GPU memory capacity, we set the volumetric size of hand pose to $80 \times 80 \times 80$. To obtain the best-performing model, we train for 8 epochs each time, requiring approximately 5 days total.

3.1 Hand Pose Datasets and Evaluation Metrics

ICVL Hand Dataset. The ICVL dataset [30] consists of a training set with 331,000 depth maps and a test set with over 1,500 depth maps, collected from 10 different hand gesture performers using an Intel Creative Interactive Gesture

Camera. Each finger in this dataset has three joints, and the palm has one joint, totaling 16 joints.

NYU Hand Dataset. The NYU dataset [31] comprises a training set with 72,000 depth maps and a test set with 8,252 depth maps. The training set images were performed by one person, while the test set was generated by two people from three different Kinect viewpoints. This dataset contains 36 joints per hand. Since most previous works only use the frontal view and 14 joints for evaluation, we also follow this configuration for fair comparison.

We evaluate hand pose estimation performance using two common metrics: the mean 3D distance error per joint (Mean Error) and the proportion of positive samples where the maximum error is below a threshold.

3.2 Experimental Results and Analysis

1) Comparison with Baseline Experiments

To investigate whether the pseudo-3D hourglass structure improves hand pose estimation, we conduct comparative experiments on the ICVL dataset [30]. In this experiment, we extend the hourglass network from 2D to 3D (3D hourglass network) as our baseline and compare it with our proposed pseudo-3D hourglass structure network. The comparison results are shown in [Figure 4: see original paper]. Additionally, to investigate whether the pseudo-3D hourglass structure reduces model size and improves estimation speed, we compare model sizes and single hand pose estimation time between our method and the baseline, with results shown in .

The experimental results show that at small error thresholds, our method's accuracy is slightly higher than the baseline, with a 3D distance mean error 0.754 mm lower than the baseline. Single hand pose estimation is 4 ms faster, and model size is significantly reduced to half of the baseline. This demonstrates that the pseudo-3D hourglass network can significantly accelerate hand pose estimation and reduce model size while slightly improving estimation accuracy.

2) Comparison with Other Methods

We compare our method with multiple approaches on both ICVL [30] and NYU [31] datasets, including Latent Random Forest (LRF) [20], Cascade [21], Deep-Prior++ [24], Hand3D [14], CrossingNets [25], Pose-REN [24], DenseReg [13], Feedback [32], and 3D CNN [1]. All results are calculated based on prediction labels provided online.

As shown in [Figure 5: see original paper], [Figure 6: see original paper], and , our method outperforms all previous methods on the ICVL dataset [30], demonstrating good performance even with small error tolerance ranges. On the NYU dataset [31], our method's accuracy is slightly lower than DenseReg [13]. This may be because the hand region is not cropped in this dataset, introducing errors during region segmentation. Nevertheless, as reported in reference [13],

DenseReg requires 36 ms for single hand pose estimation, while our method only needs 8.39 ms, making it significantly faster.

4 Conclusion

This paper proposes an accurate hand pose estimation method based on pseudo-3D convolutional neural networks. Hand depth maps are encoded into 3D volumetric representations and fed into the pseudo-3D CNN as input, using an improved segmentation method to extract hand regions. By cascading spatial and depth convolutional filters, we simplify standard 3D convolutions. Feature extraction and fusion at multiple scales enable full utilization of 3D information in hand pose. Experimental results show that our model has a small size, improves accuracy, and accelerates hand pose estimation speed. In future work, we will further investigate the impact of diverse network structures on hand pose estimation performance.

References

[1] Ge Liuhao, Liang Hui, Yuan Junsong, et al. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017.

[2] Yuan Shanxin, Ye Qi, Stenger Bjorn, et al. BigHand2. 2M benchmark: hand pose dataset and state of the art analysis [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 35-45.

[3] Shotton J, Kipman A, Blake A, et al. Efficient human pose estimation from single depth images [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2013, 35 (12): 2821-2840.

[4] Qian Chen, Sun Xiaoli, Wei Yichen, et al. Realtime and Robust Hand Tracking from Depth [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2014: 1106-1113.

[5] 徐岳峰, 周书仁, 王刚, 等. 基于深度图像梯度特征的人体姿态估计 [J]. 计算机工程, 2015, 41 (12): 200-205. (Xu Yuefeng, Zhou Shuren, Wang Gang, et al. Human body attitude estimation based on gradient feature of depth images [J]. Computer Engineering, 2015, 41 (12): 200-205.)

[6] 王松, 刘复昌, 黄骥, 等. 基于卷积神经网络的深度图姿态估计算法研究 [J]. 系统仿真学报, 2017, 29 (11): 2618-2623. (Wang Song, Liu Fuchang, Huang Ji, et al. Pose estimation using convolutional neural network with synthesis depth data [J]. Joural of System Simulation, 2017, 29 (11): 2618-2623.)

[7] Tompson J, Stein M, Lecun Y, et al. Real-time continuous pose recovery of human hands using convolutional networks [J]. ACM Trans on Graphics, 2014, 33 (5): 1-10.

[8] Oberweger M, Wohlhart P, Lepetit V. Hands deep in deep learning for hand pose estimation [EB/OL]. (2015-10-02). https://www.tugraz.at/…/3d_{{{hand}}}*{{pose}}}/cvww15*{presentat

[9] Ge Liuhao, Liang Hui, Yuan Junsong, et al. Robust 3D Hand Pose Estimation in Single Depth Images: from Single-View CNN to Multi-View CNNs [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016.

[10] Sinha A, Choi C, Ramani K. DeepHand: robust hand pose estimation by completing a matrix imputed with deep features [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 4150-4158.

[11] Zhou Xingyi, Wan Qingfu, Zhang Wei, et al. Model-based Deep Hand Pose Estimation [C]//Proc of the 25th International Joint Conference on Artificial Intelligence. 2016: 2421-2427.

[12] Tompson J, Stein M, Lecun Y, et al. Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks [J]. ACM Trans on Graphics, 2014, 33 (5): 1-10.

[13] Wan Chengde, Probst T, Van Gool L, et al. Dense 3D regression for hand pose estimation [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2018.

[14] Deng Xiaoming, Yang Shuo, Zhang Yinda, et al. Hand3D: hand pose estimation using 3D neural network [EB/OL]. (2017). http://cn. arxiv. org/pdf/1704. 02224.

[15] Qiu Zhaofan, Yao Ting, Mei Tao. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 5534-5542.

[16] Newell A, Yang Kaiyu, Deng Jia. Stacked hourglass networks for human pose estimation [C]//Proc of European Conference on Computer Vision. 2016: 483-499.

[17] Tagliasacchi A, Tkach A, Bouaziz S, et al. Robust articulated-ICP for real-time hand tracking [C]//Proc of Eurographics Symposium on Geometry Processing. 2015: 101-114.

[18] Oikonomidis I, Kyriazis N, Argyros A. Efficient model-based 3D tracking of hand articulations using Kinect [C]//Proc of British Machine Vision Conference. 2011.

[19] Shotton J, Fitzgibbon A, Cook M, et al. Real-time human pose recognition in parts from single depth images [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2011: 1297-1304.

[20] Tang Danhang, Chang Hyung Jin, Tejani A, et al. Latent regression forest: structured estimation of 3D articulated hand posture [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2014: 3786-3793.

[21] Sun Xiaoli, Wei Yichen, Liang Shuang, et al. Cascaded hand pose regression [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015.

[22] Guo Hengkai, Wang Guijin, Chen Xinghao, et al. Region ensemble network: improving convolutional network for hand pose estimation [C]//Proc of IEEE International Conference on Image Processing. 2017.

[23] Chen Xinghao, Wang Guijin, Guo Hengkai, et al. Pose guided structured region ensemble network for cascaded hand pose estimation [OL]. (2017-03). http://cn. arxiv. org/pdf/1708. 03416.

[24] Oberweger M, Lepetit V. DeepPrior+: improving fast and accurate 3D hand pose estimation [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017.

[25] Wan Chengde, Probst T, Van Gool L, et al. Crossing nets: combining GANs and VAEs with a shared latent space for hand pose estimation [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 1196-1205.

[26] Wu Zhirong, Song Shuran, Khosla A, et al. 3D ShapeNets: A deep representation for volumetric shapes [C]//IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 1912-1920.

[27] Song Shuran, Xiao Jianxiong. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 808-816.

[28] Maturana D, Scherer S. VoxNet: a 3D convolutional neural network for real-time object recognition [C]//Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems. 2015: 922-928.

[29] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep Residual Learning for Image Recognition [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 770-778.

[30] Tang Danhang, Chang Hyung Jin, Tejani A, et al. Latent regression forest: structured estimation of 3D articulated hand posture [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2014: 3786-3793.

[31] Tompson J, Stein M, Lecun Y, et al. Real-Time Continuous pose recovery of human hands using convolutional networks [J]. ACM Trans on Graphics, 2014,

33 (5): 1-10.

[32] Oberweger M, Wohlhart P, Lepetit V. Training a feedback loop for hand pose estimation [C]//Proc of International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2016.

Note: Figure translations are in progress. See original paper for figures.