# Residual Attention Regression for 3D Hand Pose Estimation

Jing Li[1[0000-0001-6566-1406]], Long Zhang[1[0000-0001-9301-6770]], Zhaojie Ju[2[0000-0002-9524-7609]]

[1] School of Information Engineering, Nanchang University, Nanchang, 330031, China
[2] School of Computing, University of Portsmouth, Portsmouth, PO1 3HE, U.K.
zhaojie.ju@port.ac.uk

**Abstract.** 3D hand pose estimation is an important and challenging task for virtual reality and human-computer interaction. In this paper, we propose a simple and effective residual attention regression model for accurate 3D hand pose estimation from a depth image. The model is trained in an end-to-end fashion. Specifically, we stack different attention modules to capture different types of attention-aware features, and then implement physical constraints of the hand by projecting the pose parameters into a lower-dimensional space. In this way, 3D coordinates of hand joints are estimated directly. The experimental results demonstrate that our proposed residual attention network can achieve superior or comparable performance on three main challenging datasets, where the average 3D error is 9.7mm on the MSRA dataset, 7.8mm on the ICVL dataset, and 17.6mm on the NYU dataset.

**Keywords:** 3D Hand Pose Estimation, Attention Mechanism, Convolutional Neural Network, Depth Images.

## 1 Introduction

Vision-based articulated hand pose estimation has made steady progress in recent years [1, 2, 3], since it is a key technology for human-computer interaction [30] in virtual reality and augmented reality applications. For the diagnosis and therapy of autism spectrum disorder (ASD), hand pose estimation can help diagnose children with autism, because the hand behaviors of ASD children are different from those of typically developing children. Moreover, it can assist ASD children to interact with robots in order to improve their social ability, i.e., robots can better understand ASD children's commands by hand pose estimation and provide appropriate feedback. Although recent progress has been achieved in 3D hand pose estimation with depth cameras [4, 5], there also remains a lot of challenges for efficient and robust estimation due to the high degree of freedom (DOF) of hand pose, severe self-occlusions, viewpoint changes, self-similarity of fingers, large variations in hand pose and background noises.

Existing human hand pose estimation methods can be categorized into two complementary paradigms: 1) learning-based (discriminative); and 2) model-based (generative). The former learns a direct regression function which maps the image appear-

ance to hand pose, and either uses random forests [6, 11, 8] or deep convolutional neural networks to estimate hand pose. Generally, it is much more efficient to evaluate the regression function than model-based optimization, but the estimation is rough and can serve as initialization for later model-based optimization [9]. Nevertheless, learning-based methods are becoming popular because they are robust and fast. The latter synthesizes images according to hand geometry, and then defines an objective function between the synthesized image and the observed image to quantify the discrepancy, and finally optimizes the objective function to obtain the hand pose.

Recently, deep convolutional neural networks (CNNs) have exhibited great performance across various computer vision tasks such as object classification [20], object detection [7, 21], semantic segmentation [22] and human pose estimation [15] because of the good modeling capacity and end-to-end feature learning. For hand pose estimation, discriminative data-driven approaches leveraging CNNs surpass traditional generative model-driven approaches in terms of accuracy and speed, and most of the recent proposed 3D hand pose estimation methods have achieved drastic performance improvement on large hand pose datasets [10, 11, 12, 13]. Thanks to the success of CNNs and the availability of low-cost depth cameras, many research efforts have been devoted to hand pose estimation from depth images, including directly taking 2D depth images as input into CNNs and outputting heat-maps [23], the 3D joint locations or the 3D pose parameters [14] such as joint angles.

Given a single depth image as input, there are two main regression-based approaches. The first approach directly regresses depth images to continuous joint 2D or 3D positions [16, 17] either in 2D or 3D space; the other approach outputs discrete heat-maps [23] for each hand joint as an intermediate result and performs some additional post-processing steps to obtain the final hand pose. However, it is non-trivial to lift 2D heat-maps to 3D joint locations. One straightforward solution is to generate volumetric heat-maps using 3D CNNs, but it is inefficient and requires much computing resources.

In this work, inspired by the attention mechanism used in image classification [19] and human pose estimation [15], we design a novel residual attention regression model based on CNNs to estimate 3D hand pose from single depth images by introducing the attention mechanism that can directly regresses the 3D joint coordinates with end-to-end training. More specifically, our main contributions are summarized as follows:

1) We propose a novel residual attention regression model to directly regress 3D coordinates of hand joints. Our network architecture is composed of multiple attention modules, generating different attention-aware features from different modules which change adaptively as layers going deeper. Our network is in an end-to-end training fashion without extra post-processing.
2) We apply a prior layer to learn a prior model from hand pose and integrate it into our network with fewer neurons on the end of the network.
3) We conduct extensive comparison experiments on three challenging hand pose datasets (e.g., MSRA datasets [12], NYU datasets [11] and ICVL datasets [10]) To evaluate the performance our proposed regression model with other representative hand pose estimation methods. Experimental results show our model is efficient and performs better on the MSRA dataset and the ICVL dataset.

## 2 Related works

3D hand pose estimation is an old and long-lasting problem in computer vision research areas. Recently, it has captured much attention due to its widespread applications in augmented reality and the popularity of depth cameras, such as Microsoft Kinect and Intel RealSense. To infer 3D hand pose, Tompson et al. [11] applied CNNs to produce 2D heat-maps which represent the probability distributions of hand joints and used model-based inverse kinematics to recover 3D hand pose from estimated heat-maps. Ge et al. [24] firstly employed 3D CNN to capture spatial features in 3D space by projecting depth images into three different views and estimating 3D hand pose from multi-view heat-maps. Oberweger et al. [16, 17] proposed a framework that directly regresses 3D coordinates of hand joints with multi-stage CNNs and used a linear layer as pose prior. They [29] also introduced a feedback loop which contains a discriminative network for initial pose estimation, a CNN for image synthesis and a CNN for refining hand pose iteratively. Chen et al. [25] designed a pose guided structured region ensemble network to capture the tree-like structure of the hand. Moon et al. [26] cast the 3D hand and human pose estimation problems from a single depth map into a voxel-to-voxel prediction which used a 3D voxelized grid and estimated the per-voxel likelihood for each keypoint of hand. Wan et al. [18] proposed to jointly train a generator for updating with the back-propagated gradient from the discriminator to synthesize realistic depth maps of the articulated hand and a discriminator to estimate the posterior of the latent pose given some depth maps. Ye et al. [27] proposed a hybrid hand pose estimation method and applied the kinematic hierarchy strategy to the input space of the discriminative method by a spatial attention mechanism in order to optimize the generative method by hierarchical Particle Swarm Optimization (PSO). Deng et al. [28] converted the depth images to a 3D volume and used a 3D CNN to predict joint locations; however, 3D networks show a low-computational efficiency.

However, most of the above-mentioned networks focus on training feedforward convolutional neural networks using a "very deep" structure to deal with the hand pose estimation problem. In this paper, we apply a residual attention network which contains a bottom-up and top-down feedforward mask branch that can generate attention-aware features to guide feature learning and regress hand joint coordinates in an end-to-end training fashion.

## 3 Model Overview

The goal of our model is to estimate $J$ 3D hand joint coordinates $\mathrm{S} = \{j_i\}_1^J$ with $j_i = (u_i, v_{i,} d_i)$ from a single depth image. Like in the previous work [16], firstly we estimate a coarse 3D bounding box containing the hand and segment the foreground based on the assumption that the hand is the closest object to the depth camera. In this way, we extract a fixed-size cube in the center of mass of this object from the depth image. Then, the cube is resized to $128 \times 128$ patch of depth values normalized to $[-1, 1]$ as the input for the CNNs. The points for which depth is not available, or the depth

values deeper than the back face of the cube, are assigned with a fixed value of 1.0. In order to be invariant to different distances between the hand and the camera for the CNN, the normalization is a key pre-processing step.

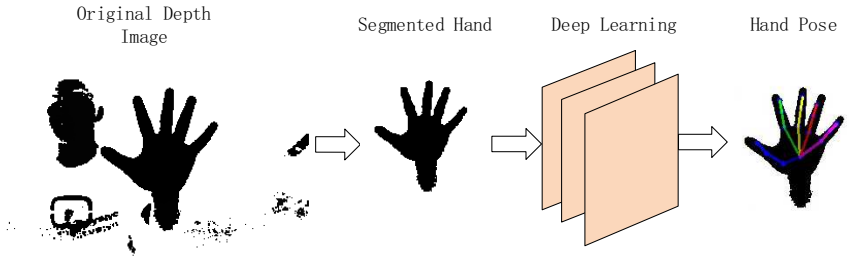


**Fig. 1.** The pipeline of hand pose estimation

The pipeline of our hand pose estimation framework is given in Fig. 1. Firstly, we segment the hand from the original depth image, and then we feed it into our network as input to generate the coordinates of hand joints which are used to recover the hand pose. The architecture of our proposed residual attention regression model is shown in Fig. 2. It is constructed by stacking three attention modules with a residual unit between two modules. The network accepts a 128×128 depth image as the input. Each attention module consists of two branches: the trunk branch and the mask branch. The trunk branch is for feature extraction and we use residual units as basic units of our residual attention module. Denote the trunk branch output $T(x)$ with input $x$, the mask branch has the same size as $M(x)$ which uses a bottom-up and top-down structure. The output of the attention module $H$ is:

$$H_{i,c}(x) = \left(1 + M_{i,c}(x)\right) * T_{i,c}(x) \tag{1}$$

where $M_{i,c}(x)$ ranges from [0, 1] for it is after a sigmoid layer, $T_{i,c}(x)$ indicates the features learned by the deep convolutional neural network. The mask branches are the key for attention modules since they have the ability to enhance good features and suppress noises from trunk features as feature selectors.

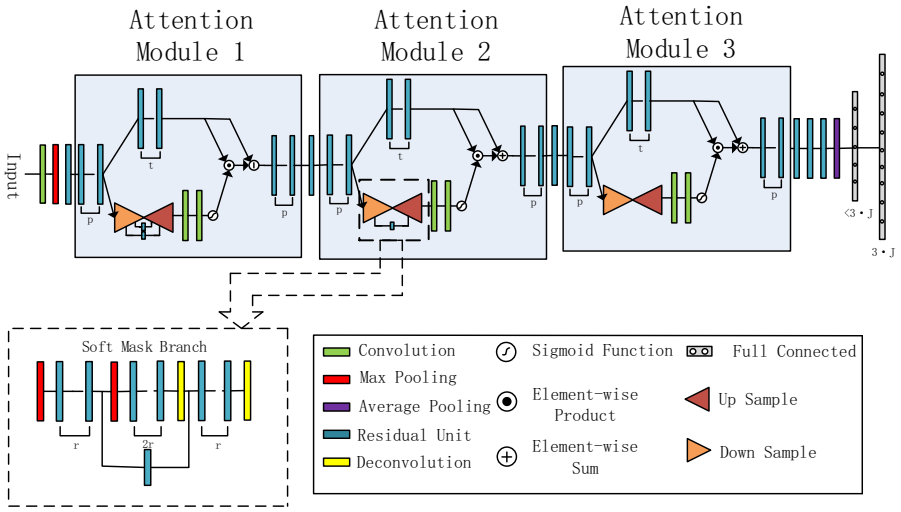


**Fig. 2.** The architecture of our proposed residual attention regression model. It consists of three different modules and there are three hyper-parameters: p, t, and r. The last fully-connected layer is the prior layer.

At the end of the network, instead of directly predicting the 3D joint locations, we predict the pose parameters in a lower-dimensional space. A previous work [16] demonstrated that this can improve the prediction reliability because it enforces constraints of the hand pose. We apply Principal Component Analysis (PCA) to learn a low-dimensional representation of the training data, which implements the physical constraints of the hand. Then, we compute the hand pose parameters by projecting them into the 3·J-dimensional joint space with the last prior layer. In order to show the estimated hand pose in the depth image, we project the predicted real-world 3D coordinates into the image pixel coordinates using the intrinsic parameters of the depth camera, as shown in Eq. 2:

$$(P_{x,i}, P_{y,i}, P_{z,i}) = proj\ (P_{u,i}, P_{v,i}, P_{d,i}) \tag{2}$$

where $(P_{x,i}, P_{y,i})$ is in pixel coordinate and $P_{z,i}$ is the distance between the depth camera and the object.

## 4 Experiment

We evaluate our proposed deep convolutional network architecture on three public hand pose datasets: MSRA Hand Pose Dataset [12], ICVL Hand Pose Dataset [10] and NYU Hand Pose Dataset [11].

**The MSRA dataset** [12] contains over 76k depth frames captured by Intel's Creative Interactive Gesture Camera which uses a time-of-flight. For each frame, it contains 3D locations of 21 annotated hand joints. The dataset contains 9 subjects. For

each subject, there are 17 gestures, each of which contains about 500 segmented hand images from depth images. We train our neural network on 8 subjects and evaluate on the remaining one.

**The ICVL dataset** [10] contains about 330k training samples and 1.6k testing samples and there is a large discrepancy between the training and testing sequences. The dataset was recorded using a time-of-flight Intel Creative Interactive Gesture Camera and the ground truth of each frame contains 16 hand joint locations with (x, y) in pixels and z in mm, and thus we should project the image pixel coordinates into the real-world 3D space by using the intrinsic parameters of the depth camera. Although different artificially in-plane rotations were applied to the collected samples, we only use the original 22k samples.

**The NYU dataset** [11] contains about 72k depth images for training and 8k depth images for testing. Each frame was captured using the Primesense Carmine 1.09, which is a structured light-based sensor and the depth maps show missing values are mostly along the occluding boundaries as well as noisy outlines. The ground truth annotation of each depth image contains 3D locations of 36 hand joints, but we just use a subset of 14 hand joints as in previous works [16, 17, 18, 28]. The dataset contains very noisy images and has a very wide range of poses, making a challenge to most hand pose estimation methods.

We adopt two different evaluation metrics to evaluate the performance of our proposed hand pose estimation method. The first metric is the per-joint mean error distance between the predicted 3D joint location and the ground truth overall testing frames as well as the overall mean error distance for all joints on all testing frames, as shown in Eq. 3. As given by Eq. 4, the second metric is the percentage of good frames in which the worst joint error is below a given threshold, which is more challenging and strict. We remove the mean error distance of the center joint for it is used for normalization.

$$avg\ error = \frac{\sum_{s=1}^{S} \sum_{j=1}^{J} Err_{(s,j)}}{S*J} \tag{3}$$

where $Err_{(s,j)}$ is the Euclidean Distance between the ground truth and the predicted location of each joint, S is the number of testing frames, and J is the number of joints in each frame.

$$per = \frac{n}{S} \tag{4}$$

where n indicates the number of frames in which the average joint error is within a maximum distance from the ground truth, and S is the number of testing frames.

We train and evaluate our proposed deep neural network model on a computer equipped with an Intel Core I7 CPU, 32GB of RAM, and two GeForce GTX 1080Ti GPUs. Our deep neural network model is implemented in Python within the TensorFlow framework. We use Adam optimizer with initial learning rate 0.0001, batch size 128, L2 regularization weight delay rate 0.0003, and the learning rate is divided by 10 after 30 epochs. We stop the training process after 70 epochs to avoid overfitting. We do not perform any data augmentation on these datasets.

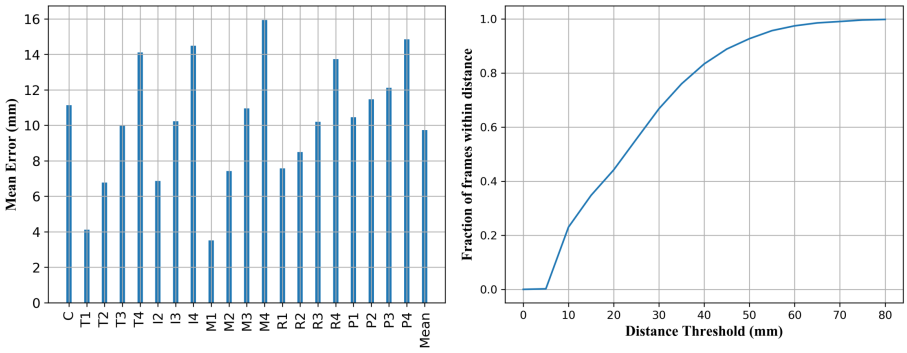


**Fig. 3.** Distance error (left) and percentage of success frames (right) on the MSRA dataset.

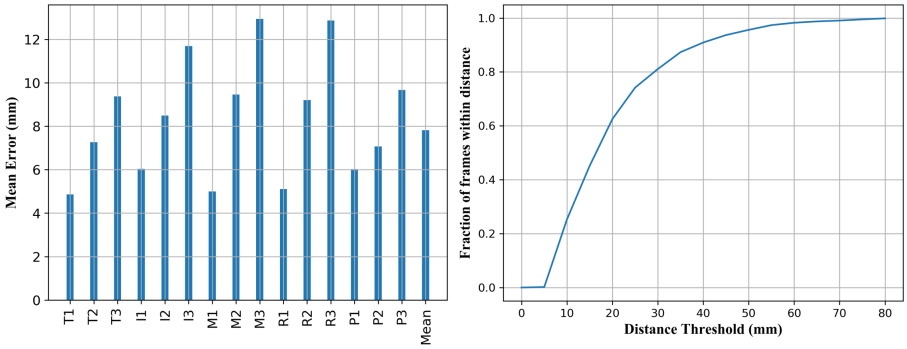


**Fig. 4.** Distance error (left) and percentage of success frames (right) on the ICVL dataset.

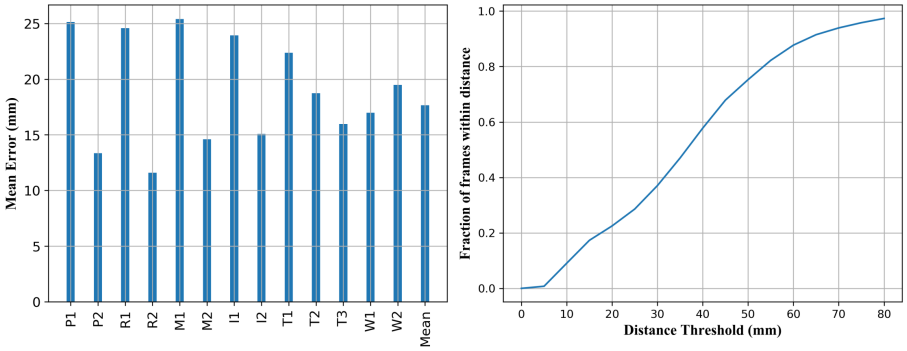


**Fig. 5.** Distance error (left) and percentage of success frames (right) on the NYU dataset.

In Fig. 3, Fig. 4 and Fig. 5, we give the results of our model on the three datasets mentioned above by adopting two different evaluation metrics. The average 3D errors are 9.7mm, 7.8mm, 17.6mm on MSRA dataset, ICVL dataset and NYU dataset, re-

8

spectively. And we also compare the 3D average error of our methods with several popular approaches in Table 1. We can see that our approach outperforms Wan et al. [18] and Oberweger et al. [16] with about 20% and 13% absolute improvement on the NYU dataset and ICVL dataset, respectively. Fig. 6 shows qualitative hand pose estimation results on three datasets.

**Table 1.** The 3D average error comparison of different methods on MSRA, ICVL and NYU datasets without augmentation, the "aug" denotes data augmentation.

| Method \ Dataset | MSRA | ICVL | NYU |
|---|---|---|---|
| Oberweger et al. [16] | — | 9.0 mm | 20.7 mm |
| Wan et al. [18] | 12.2 mm | 10.2 mm | **15.5 mm** |
| Sun et al. [12] | 15.2mm | — | — |
| Deng et al. [28] | — | 10 - 11 mm (aug) | 17.6 mm (aug) |
| **Our** | **9.7 mm** | **7.8 mm** | 17.6 mm |

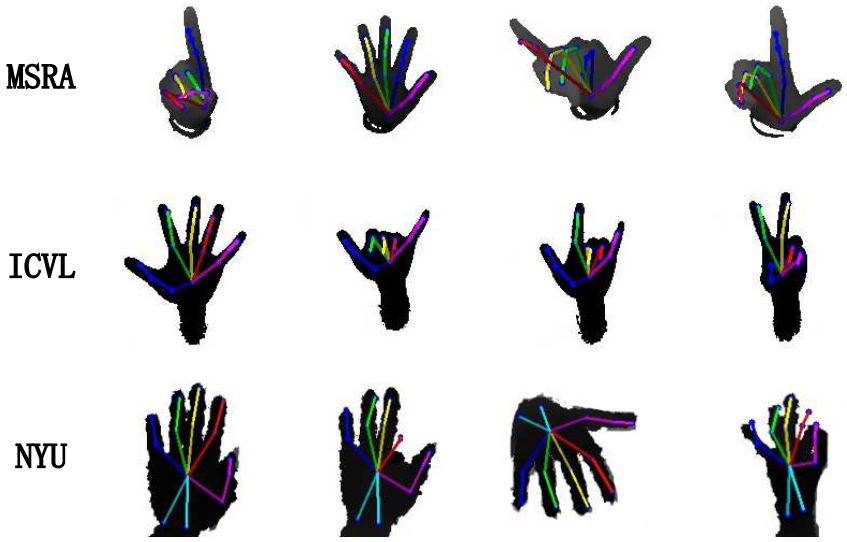


**Fig. 6.** Qualitative hand pose estimation results on three datasets.

## 5 Conclusions

In this paper, we propose a novel 3D hand pose regression model based on residual attention network from a single depth image. Our method applies different attention modules to capture different types of attention to guide feature learning for 3D hand pose regression and can predict accurate real-world 3D coordinates of hand joints. To avoid the distance influence between the camera and the hand, we crop the hand with a fixed cube according to the center of mass and normalize the depth image and the ground truth. Further, we apply pose prior to add the physical constraints of the hand to estimate accurate hand pose. Experimental results on the three challenging hand pose datasets show that our model achieves superior or comparable performance without data augmentation.

## References

1. Hui, L., Yuan, J., Thalmann, D.: Resolving ambiguous hand pose predictions by exploiting part correlations. IEEE Transactions on Circuits and Systems for Video Technology 25(7), 1125–1139 (2015).
2. Oberweger, M., Riegler G., Wohlhart, P., Vincent, L.: Efficiently creating 3d training data for fine hand pose estimation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4957-4965, IEEE, Las Vegas, NV, USA (2016).
3. Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3D tracking of hand articulations using Kinect. In: Proceedings of the 22nd British Machine Vision Conference (BMVC), pp. 101.1--101.11, BMVA Press, University of Dundee, UK (2011).
4. Xu, C., Cheng, L.: Efficient hand pose estimation from a single depth image. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 3456-3462, IEEE, Sydney, NSW, Australia (2013).
5. Ge, L., Liang, H., Yuan, J., Thalmann, D.: 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5679-5688, IEEE, S Honolulu, HI, USA (2017).
6. Tang, D., Yu, T., Kim, T.K.: Real-time articulated hand pose estimation using semisupervised transductive regression forests. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 3224–3231, IEEE, Sydney, NSW, Australia (2013).
7. Liu, W., Anguelov, D., Erhan, D., Szegedy, C.: SSD: Single shot multibox detector. In: 2016 European Conference on Computer Vision (ECCV), pp. 21–37, Springer Cham, Amsterdam, The Netherlands (2016).
8. Li, P., Ling, H., Li, X., Liao, C.: 3d hand pose estimation using randomized decision forest with segmentation index points. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 819–827, IEEE, Santiago, Chile (2015).
9. Srinath, S., Franziska, M., Antti O., Christian T.: Fast and robust hand tracking using detection-guided optimization. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3221, IEEE, Boston, MA, USA (2015).
10. Tang, D., Chang, H.J., Tejani, A., Kim, T.K.: Latent regression forest: Structured estimation of 3D articulated hand posture. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3786–3793, IEEE, Columbus, OH, USA (2014).

11. Tompson, J., Stein, M., Lecun, Y., Perlin, K.: Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics 33(5), 169 (2014).
12. Sun, X., Wei, Y., Liang, S., Tang, X., Sun, J.: Cascaded hand pose regression. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 824–832, IEEE, Boston, MA, USA (2015).
13. Yuan, S., Ye, Q., Stenger, B., Jain, S., Kim, T.K.: Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2605–2613, IEEE, Honolulu, HI, USA (2017).
14. Zhou, X., Wan, Q., Zhang, W., Xue, X., Wei, Y.: Model based deep hand pose estimation. Proceeding IJCAI'16 Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2421-2427 (2016).
15. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: 2016 European Conference on Computer Vision (ECCV), pp. 483–499, Springer Cham, Am-sterdam, The Netherlands (2016).
16. Oberweger, M., Wohlhart, P., Lepetit, V.: Hands deep in deep learning for hand pose estimation. In: 2015 Computer Vision Winter Workshop (CVWW), pp. 1-10 (2015).
17. Oberweger, M., Lepetit, V.: Deepprior++: Improving fast and accurate 3d hand pose estimation. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) (2017), pp. 585-594, IEEE, Venice, Italy (2017).
18. Wan, C., Thomas P., Van Gool, L., Yao, A.: Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1196-1205, IEEE, Honolulu, HI, USA (2017).
19. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H.: Residual attention network for image classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6450-6458, IEEE, Honolulu, HI, USA (2017).
20. Alex, K., Ilya S., Geoffrey E Hinton.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25(2), 1097–1105 (2012).
21. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence 38(1), 142–158 (2016).
22. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv:1606.00915 (2016).
23. Choi, C., Kim, S., Ramani, K.: Learning hand articulations by hallucinating heat distribution. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017), pp. 3123-3132, IEEE, Venice, Italy (2017).
24. Ge, L., Liang, H., Yuan, J., Thalmann, D.: Robust 3d hand pose estimation from single depth images using multi-view cnns. IEEE Trans. Image Processing 27(9), 4422-4436 (2018).
25. Chen, X., Wang, G., Guo, H., Zhang, C.: Pose guided structured region ensemble network for cascaded hand pose estimation. arXiv:1708.03416 (2017).
26. Moon, G., Chang, J.Y., Lee, K.M.: V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5079–5088, IEEE, Salt Lake City, UT, USA (2018).

27. Ye, Q., Yuan, S., Kim, T.-K.: Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In: 2016 European Conference on Computer Vision (ECCV), pp. 346–361, Springer Cham, Amsterdam, The Netherlands (2016).
28. Deng, X., Yang, S., Zhang, Y., Tan, P., Chang, L., Wang, H.: Hand3d: Hand pose estimation using 3d neural network. arXiv:1704.02224 (2017).
29. Oberweger, M., Wohlhart, P., Lepetit, V.: Training a feedback loop for hand pose estimation. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3316–3324, IEEE, Santiago, Chile (2015).
30. Li, J., Wang, J., Ju, Z.: A novel hand gesture recognition based on high-level features. International Journal of Humanoid Robotics 15(1), 1750022 (2018).