Structure-Aware 3D Hourglass Network for Hand Pose Estimation from Single Depth Image

Fuyang Huang[1]
fyhuang@cse.cuhk.edu.hk

Ailing Zeng[1]
alzeng@cse.cuhk.edu.hk

Minhao Liu[1]
mhliu@cse.cuhk.edu.hk

Jing Qin[2]
harry.qin@polyu.edu.hk

Qiang Xu[1]
qxu@cse.cuhk.edu.hk

[1] Computer Science & Engineering
The Chinese University of Hong Kong
Hong Kong

[2] School of Nursing
The Hong Kong Polytechnic University
Hong Kong

Abstract

In this paper, we propose a novel structure-aware 3D hourglass network for hand pose estimation from a single depth image, which achieves state-of-the-art results on MSRA and NYU datasets. Compared to existing works that perform image-to-coordination regression, our network takes 3D voxel as input and directly regresses 3D heatmap for each joint. To be specific, we use hourglass network as our backbone network and modify it into 3D form. We explicitly model tree-like finger bone into the network as well as in the loss function in an end-to-end manner, in order to take the skeleton constraints into consideration. Final estimation can then be easily obtained from voxel density map with simple post-processing. Experimental results show that the proposed structure-aware 3D hourglass network is able to achieve a mean joint error of 7.4 mm in MSRA and 8.9 mm in NYU datasets, respectively.

1 Introduction

Accurate gesture recognition plays an important role in many applications (e.g., virtual reality and augmented reality). Articulated hand pose estimation, serving as a fundamental step towards gesture recognition, has thus drawn great attention from both industry and academia.

Thanks to the availability of low-cost depth camera and recent advancements in machine learning techniques, many research efforts have been devoted to hand pose estimation from depth image in recent years. Despite the significant improvement in terms of accuracy and speed, current articulated hand pose estimation solution is still far from satisfactory. The main challenge lies in the fact that regressing from depth image to 3D coordination is a highly non-linear problem, which involves high degree of freedom (DOF) of hand pose, frequent self-occlusion, and background noises.

In recent years, convolutional neural network (CNN) has been successfully applied in various types of computer vision tasks such as object detection, image classification, and human pose estimation [5, 18, 28]. Similarly, discriminative data-driven approaches leveraging deep learning method overwhelm traditional generative model-driven approaches for hand pose estimation tasks in terms of accuracy and speed. It is regarded as a non-linear regression problem which regresses depth image to 3D joint coordination [20, 25] and trained in a deep learning network.

There are two main kinds of regression-based approaches. One directly regresses depth image to continuous joint positions [4, 11, 12, 19, 20] either in 2D or 3D space. The other outputs discrete probability heatmaps for each joint as a intermediate result and requires additional post-processing to get the final result. Recently, the heatmap-based approach has proved to be more promising in both human pose estimation [1, 5, 9, 15, 24, 28, 29] and hand pose estimation [10, 16, 25, 27]. Wan et al. [27] proposed a network that produces 2D heatmaps and 3D offset vector heatmaps to get the final 3D joint positions. It takes 2D depth image as input and employs 2D CNN to regress 3D heatmap. The depth data is originally captured in 3D space, but most of 2D CNN-based work uses projected 2D depth image to regress 3D coordination. Information loss is inevitable during the 3D-to-2D projection and 2D-to-3D regression. Moon et al. [16] proposed a voxel-to-voxel network with 3D ResNet [14] as backbone network. They directly regress 3D heatmap of joint from 3D voxel to preserve as much information as possible,but they do not take skeleton constraints of hand pose into consideration in this network, which is essential for accurate hand pose estimation [4, 19, 25].

To tackle the above problems, we propose a novel structure-aware 3D hourglass network in this paper. Our framework is able to capture 3D spatial feature of depth data in all scale and embed skeleton constraints of hand pose in a single network. The input depth data and ground truth are first transformed into camera's coordination system and then discretized to voxels in 3D space. Subsequently, the voxels are regressed to 3D heatmaps of each joint, from which the final joint positions can be easily retrieved with simple post-processing. The regression model is trained by a stacked 3D hourglass network that captures all scales of 3D convolutional features in an end-to-end manner. In order to take skeleton constraints into consideration, we follow the tree structure of human hand and add additional channels at the final layer of each stack as bone heatmaps. These bone heatmaps work as intermediate supervision in the loss function and are then passed to next stack of hourglass network as input. Note that, the non-linear relationship between joint and skeleton structure can be learned implicitly in the following stacks of networks. Accordingly, the final joint heatmaps will be refined by the skeleton constraints of hand pose. To the best of our knowledge, this is the first work that regresses 3D depth data into 3D heatmaps with skeleton constraints for 3D hand pose estimation. Experiments on two popular public datasets (MSRA [22] and NYU [25]) show that our method outperforms most of the state-of-the-art works in terms of mean joint error and success frame rate. To sum up, the main contributions of this work include:

- This is the first work that incorporates hourglass network with depth image in purely 3D form to regress 3D joint heatmaps directly. Unlike most of previous hand pose estimation work leveraging hourglass network, our network consists of 3D residual network (ResNet) as basic building blocks and directly outputs 3D heatmap for each joint.

- This is the first work that incorporates skeleton constraints of hand pose into de-

tection network as intermediate supervision, which is different from most heatmap-based work that concatenates another refinement network or applies complex post-processing. This method is experimentally proved to be simple yet effective in hand pose estimation.

The remainder of this paper is organized as follows. In Section 2, literature on hand pose estimation is reviewed. In Section 3, architecture of our method is introduced. We show ablation study and comparison study with state-of-the-art work in Section 4. Section 5 concludes this paper.

2 Related Work

Generative Approach vs Discriminative Approach Hand pose estimation from depth image has been extensively explored in recent years [12, 16, 17, 19, 21, 23, 26, 27]. There are two main streams of methods for this task, namely, model-driven generative approach and data-driven discriminative approach. The generative approach typically fits a pre-defined 3D hand model to a specific pose to match the input depth image [2, 21], which can be regarded as an optimization problem. By shrinking the searching space, temporal information is usually involved in these optimizers such as Particle Swarm Optimizer (PSO) [21, 22] and Iterative Closest Point (ICP) [23]. Despite high accuracy of generative methods, these methods highly rely on model initialization and temporal information, which lead to accumulative estimation error and re-initialization process in the running time.

The discriminative approach has drawn more research attention these days [2, 3, 6, 7, 8, 13, 27] to directly localize hand joints from an input depth map. Accordingly, the literature regards the depth data either as a 2D depth image [10, 12, 19, 25, 27] or as a 3D point cloud [11, 16] in the pre-processing stage. Based on the data representation, the network structure of regression model can be categorized into 2D CNN-based and 3D CNN-based respectively. Considering the high non-linear relationship between input and output 3D joint positions, instead of regressing 3D joint position directly [4, 19, 20], some of the researchers alternatively use heatmap to estimate final result [10, 16, 25, 27]

Direct Regression vs Joint Heatmap For the direct regression model, a fully connected (FC) layer is appended to the end of the network to output continuous joint positions directly. Oberweger and Lepetit [19] propose a framework introducing a bottleneck layer at the end of the network to direct regress 3D joint position while preserving skeleton constraints in the network. Chen et al. [4] propose a pose guided structured region ensemble network which is designed to capture tree-like structure of hand and is trained end-to-end in an iterative way. Tompson et al. [25] is the first work that employs CNN to generate the joint heatmap. However, post-processing on 2D heatmap has an important limitation that self-occlusion of hand joints because they may share same depth value. To provide more data in different perspective, Ge et al. [10] project depth image into three different views and apply multi-view CNN to get multi-view heatmaps. Nevertheless, re-projecting 2D image to 3D space still underutilized spatial information of depth data.

2D CNN vs 3D CNN Wan et al. [27] use 2D CNN regress 2D heatmap and 3D heatmap at the same time, although there is information loss in the 2D-to-3D regression. Ge et al. [11] firstly employs 3D CNN to capture spatial feature in 3D space. Moon et al. [16] discretize 3D point cloud data into voxel grids and push them into a voxel-to-voxel network, which produces a set of 3D heatmaps for each joint. The network structure applies 3D ResNet as
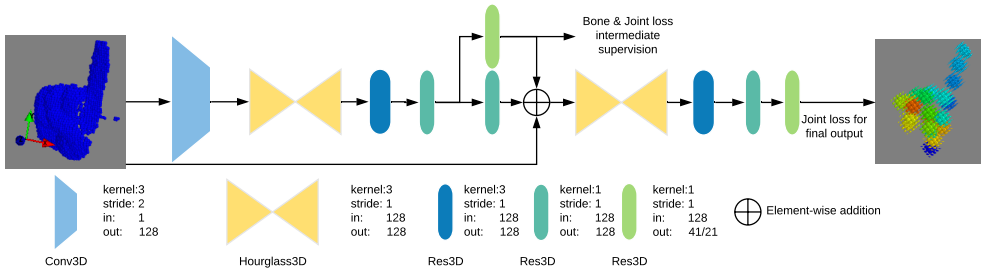
Figure 1: To simplify the illustration, all the 3D modules are visualized with 2D shapes. The two-stacks 3D hourglass network starts with discretized 3D volumetric data. The volume data is first down-sampled to $32 \times 32 \times 32$ by a 3×3 3D CNN layer with stride 2. After feature maps propagates through the first hourglass module and Res3D module, bone and joint heatmaps are generated by two consecutive 1×1 Res3D modules as intermediate supervision. The heatmaps, together with original data and feature maps, are fed into next stack for further propagation.

basic building block which is organized in a down-to-up structure and achieves state-of-the-art results in three public datasets. However, skeleton constraints of hand pose is not taken into consideration in this network which proved important in hand pose estimation system.

According to aforementioned facts, we argue that 3D CNN is more suitable for 2.5D depth data because there is no information loss in the 3D-to-3D regression compared to 2D-to-3D regression. Meanwhile, 3D data presentation better preserves spatial information of depth data compared to the 2D image. Although more 3D CNN-based work has emerged recently, the power of 3D CNN is not fully investigated due to shallow network structure [11] or skeleton-unawareness [16]. So we propose a structure-aware 3D hourglass network to estimate hand pose. Our network takes 3D voxels as input and outputs 3D heatmap for each joint, employing 3D hourglass network as basic building block. Instead of performing IK-like optimization as post-processing, skeleton constraints of hand pose is treated as an intermediate supervision in the network and trained in an end-to-end manner.

3 Proposed Method

We define the 3D hand pose estimation from depth image as a voxel-wise regression problem. Our method takes normalized binary voxels as input and outputs 3D heatmap of hand joints in voxel space. The input depth image is first reprojected to camera space as point cloud data. Then, the point cloud data is discretized into binary voxels where normalization is performed according to the size of point cloud and then the binary voxels are passed to the network for forward propagation. The details of the proposed network is described in section 3.2. The training target is 3D Gaussian distribution of joint probability in voxel form generated from ground truth. To involve in the skeleton constraints of hand pose, 3D bone probability heatmaps are trained as an intermediate supervision in the network simultaneously. The overall framework is shown in Fig 1.

3.1 Pre-processing

Before being fed into a 3D hourglass network, input data should be transformed in volumetric representation. The depth images are first re-projected to camera space by camera's intrinsic parameter. In order to truncate the point cloud into a cubic box, we simply regard the geometric center of points as box center and maximum length along x-,y-, and z-axis as box length. Afterward, the point cloud will be discretized into binary voxel grids where voxel with value one indicating it is occupied by depth data and zero otherwise. As for the ground truth, they are also transformed into voxel grids. However, the ground truth heatmap for one joint only contains one positive point which is too sparse for training. So we perform 3D Gaussian distribution on the ground truth voxels whose center is located at the ground truth point with the standard deviation of voxel grid length. A sample heatmap image is also shown in Fig 1.

3.2 Network Structure

The basic building block of our network is hourglass network proposed by Newell et al. [13] in 2016 which is widely used in human pose estimation. The bottom-up and top-down scheme of hourglass network is effective in dense prediction tasks such as semantic segmentation and heatmap prediction. Additionally, a set of hourglass networks can be stacked together to employ intermediate supervision where we can involve in skeleton constraints of hand.

Different from [13], our network takes 3D volumetric data as input. Thus, 3D CNN layers consist the basic module of the network. Considering the large consumption of GPU memory of 3D CNN, we set the input resolution of voxel to $64 \times 64 \times 64$, output resolution to $32 \times 32 \times 32$ and stack number to 2. Before being fed into hourglass module, the input voxels are down-sampled by a $3 \times 3 \times 3$ CNN layer with stride 2. The number of channels of each residual block in hourglass is 128. After forwarding through each hourglass module, the voxel-wise heatmaps of J joints are predicted by two consecutive $1 \times 1 \times 1$ CNN layers at the end of each stack.

3.3 Training Target

To predict voxel-wise heatmap for each joint, we need to generate training target first. Since there is only one ground truth value for each joint in a single 3D heatmap, direct regression may lead to overfitting. Similar to many dense prediction work before [13, 25], we perform 3D Gaussian distribution centered at the ground truth position of each joint with standard deviation of voxel length, which is then compared to the predicted heatmap by Mean-Squared Error (MSE) loss. Specially, the loss for joints is defined in equation 1:

$$L_j = \sum_{s=1}^{S} \sum_{n=1}^{J} \sum_{i,j,k}^{R} |H_{s,n}^j(i,j,k) - \hat{H}_n^j(i,j,k)|^2 \qquad (1)$$

where $H_{s,n}^j(i,j,k)$ indicates the predicted value at i,j,k of heatmap for nth joint in sth stack. $\hat{H}_n^j$ is the ground truth heatmap for nth joint. S, J and R denote stack number, joint number and resolution respectively.

3.4 Skeleton constraints

As discussed before, explicitly leveraging skeleton constraints of hand helps to increase prediction accuracy in hand pose estimation task [4, 19]. However, heatmap-based hand pose estimation tasks either let the skeleton constraints trained implicitly by the network [16, 22] or define it as an additional optimization problem in the post processing stage [10]. In our proposed network, besides heatmap for each joint, heatmaps for each bone are also learned in the network as an intermediate supervision. The number of bones are determined by the tree structure of human hand. For example, in MSRA dataset, there are 21 joints in each sample, connected by 20 bones. The bone heatmaps are generated in the similar way as joint heatmap with smaller standard deviation which equals to 0.5 times of voxel length. Additionally, bone-to-joint relationship can be learned in our stacked network, benefiting joint refinement in the final stack. In order to learn the bone-to-joint relationship, bone heatmaps are supervised at the end of each stack except for the last stack, as shown in Fig 1. The bone loss as intermediate supervision is defined in equation 2.

$$L_b = \sum_{s=1}^{S-1} \sum_{n=1}^{B} \sum_{i,j,k}^{R} |H_{s,n}^b(i,j,k) - \hat{H}_n^b(i,j,k)|^2 \tag{2}$$

where $H_{s,n}^b(i,j,k)$ represents the predicted value at i,j,k of heatmap for nth bone in sth stack and B denotes the total number of bones.

Thus, the final loss function for the entire network is defined as $L = L_j + L_b$

3.5 Post processing

Because the average length of one voxel with $32 \times 32 \times 32$ resolution is 8mm and if we simply choose the top responding voxel as final result, the error caused by discretization could be as much as $\sqrt{4^2 + 4^2 + 4^2} = 6.92$mm. Hence, we adopt a simple strategy to recover joint position as shown in equation 3:

$$J_n = \sum_{i=1}^{K} w_i^{(n)} j_i^{(n)}$$
$$\text{where } w_i^{(n)} = \frac{H_n^j(j_i^{(n)})}{\sum_{i=1}^{K} H_n^j(j_i^{(n)})} \tag{3}$$

For nth joint, the final position is the weighted mean of top K responding voxel of heatmap n. $j_i^{(n)}$ denotes the position of top i responding voxel for joint n and $w_i^{(n)}$ denotes the weight of each voxel whose sum is normalize to 1 by the corresponding value $H_n^j(j_i^{(n)})$ of predicted joint heatmap n. This strategy contributes 0.3mm accuracy to final result in MSRA dataset.

4 Experimental Results

Our method is tested on two public datasets —MSRA [22] and NYU [25]. We measure the mean error distance in 3D space and percentage of success frame (error of all joints in a single frame are below a threshold) as our measuring metrics. Hand region is cropped out in MSRA dataset, while NYU dataset retains the original depth data including background

which requires additional hand detection step. To test our hand pose estimation network in a fair way, we use MSRA dataset in the ablation study to eliminate the influence of hand detection.

4.1 Datasets

MSRA Hand Pose Dataset. There are 75K images in MSRA dataset consisting of 17 gestures from 9 subjects [22]. Hand region is cropped out in this dataset, but testing set is not specified. Similar to what most previous work do [16, 26], we apply leave-one-out cross-validation strategy to split the dataset into training set with 8 subjects and testing set with 1 subject.

NYU Hand Pose Dataset. There are 72k images for training and 8k images for testing in NYU dataset [25] whose annotations of hand pose contain 36 joints. Because hand region is not cropped out in this dataset, we directly crop the hand region in a cubic box whose center is located at center-of-mass of hand from ground truth. In our experiment, we also followed most of the previous works using frames from the frontal view and 14 out of 36 joints in the evaluation.

4.2 Experiment Setting

In order to facilitate different orientation and aspect ratio of input data, we perform data augmentation, which proved to be an effective approach in discriminative approach-based hand pose estimation task [19]. Specifically, we implement 3D data augmentation by rotating point cloud and by changing the aspect ratio in xy space of the 3D space. Each training sample is randomly rotated from $-30°$ to $30°$ and the aspect ratio is scaled from 0.8 to 1.2 during training.

We run RMSProp optimizer in the training stage. All weights are initialized randomly from scratch. The input and output resolution is $64 \times 64 \times 64$ and $32 \times 32 \times 32$ respectively. The learning rate is initially set to 1e-5 and decays 0.3 every 5 epochs. Our system is implemented by PyTorch and trained on a single NVIDIA TITAN X GPU with batch size 16 and epoch number 20.

4.3 Ablation Study

In the ablation study, we aim to answer three questions: Does heatmap-based regression outperforms direct coordination regression? Does 3D data representation better interprets 2.5D depth data over 2D representation in hand pose estimation task? Does skeleton constraints term help to improve accuracy?

Direct Regression & Heatmap. For the first experiment, we want to find whether heatmap-based regression outperforms direct coordination direction. So the first baseline shares the same network basic building block (3D hourglass) with the our proposed network except for some changes in the end of network. For baseline 1 direct regression, two consecutive fully connected layers replace original 1×1 convolutional layer to direct output 3D joint positions without skeleton constraint loss. For baseline 2, we only remove the skeleton constraints in the output layer compared to our proposed network. By comparing the result of baseline 1 and 2, direct coordination regression performs inferior to heatmap-based method with regard to mean joint error as shown in Table 1.

Baselines	Mean Joint Error
3D Direct regression without skeleton loss (B1)	11.6mm
3D Heatmap without skeleton loss (B2)	7.9mm
2D Direct regression without skeleton loss (B3)	13.1mm
2D heatmap without skeleton loss (B4)	8.1*mm
2D heatmap with skeleton loss (B5)	7.6*mm
3D heatmap with skeleton loss (Proposed)	7.4mm

Table 1: Mean joint error results for five baselines and proposed method on MSRA dataset. B is short for baseline. $*$ indicates that error is only measured in xy space.
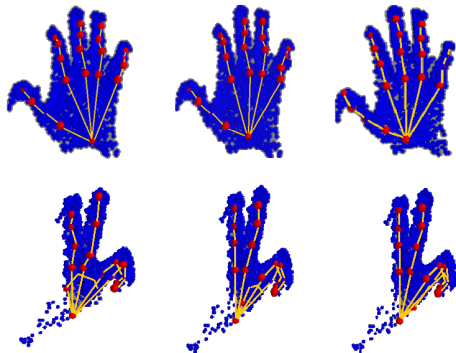

Figure 2: Qualitative results of Baseline 2 (left), proposed method (middle) and ground truth (right) on MSRA dataset, visualized in voxel space.

2D & 3D Data Representation.   For the second experiment, we evaluate the impact of 2D and 3D data representation on estimation. Here we introduce baseline 3 which takes 2D depth image as input and consists of 2D hourglass modules. The output layer is the same as that of baseline 1 network which regresses 3D coordination directly. As can be seen in Table 1, 3D data representation outperforms 2D representation by a large margin. Meanwhile, it is more compute-intensive.because of 3D convolutional operation.

Skeleton constraints.   For the last experiment, we evaluate the effectiveness of bone loss in our network quantitatively and qualitatively. For quantitative comparison, a straight way is to compare the result of baseline 2 and proposed method which indicates that skeleton constraints reduces the mean error by 1mm. To further prove the effectiveness of bone loss, we conduct baseline 4 and baseline 5 which apply 2D hourglass network and output 2D heatmap. Baseline 5 involves in bone heatmap while baseline 4 does not. Because it is 2D heatmap, we only measure the joint error in xy space. The performance of baseline 5 is also better than baseline 4. We additionally conduct qualitative comparison between baseline 2 and proposed method. As shown in Fig 2, compared to baseline 2, result of proposed method is more realistic. To be specific, as shown in the second row, finger bones are constrained by skeleton model, showing a more straight bone and a more reasonable bone length than baseline 2 result. We can conclude that the bone heatmap helps in hand pose estimation.

Method (MSRA)	Mean Error
Multiview-CNN [10]	13.2mm
Deepprior++ [19]	9.5mm
3DCNN [11]	9.5mm
Crossing Net [26]	12.2mm
Cascaded [22]	15.2mm
REN [12]	9.7mm
Pose-REN [4]	8.65mm
V2V-net [16]	7.59mm
Dense [27]	**7.2mm**
3D hourglass (Ours)	7.4mm

Method (NYU)	Mean Error
Deepprior++ [19]	12.24mm
3DCNN [11]	14.1mm
Crossing Net [26]	15.5mm
V2V-net [16]	**8.42mm**
REN [12]	12.69mm
Pose-REN [4]	11.81mm
Dense [27]	10.2mm
3D hourglass (Ours)	8.9mm

Table 2: Comparison result regarding mean joint error on MSRA and NYU dataset.
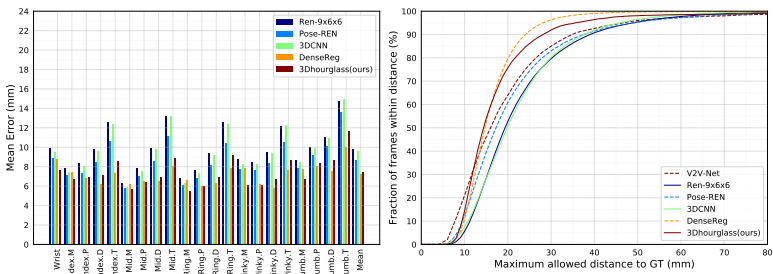

Figure 3: Result of per-joint mean error (left) and success frame rate (right) on MSRA dataset

4.4 Comparison to State-of-the-art Solutions

We compare our proposed network to several state-of-the-art hand pose estimation work on MSRA and NYU dataset [4, 10, 11, 12, 16, 19, 22, 26, 27]. Specifically, V2V-net [16] and 3DCNN [11] are 3D CNN-based work. Cascaded regression (Cascaded) [22], region ensemble network (REN) [12], pose-guided regional ensemble network (Pose-REN) [4], Deepprior++ [19] and GAN-based Crossing Net [26] are direct regression work. Dense regression model (DENSE) [27] and Multiview-CNN [10] are heatmap-based work.

As shown in Fig 3 and Table 2, our work outperforms most of the state-of-the-art works in terms of mean joint error and success frame. For mean joint error, our result ranks second in both of MSRA dataset (7.4mm) and NYU dataset (8.9mm) separately, indicating its advantage of cross-dataset stability. For the success frame, our work achieves the best result on MSRA dataset when threshold is around 15mm and the second best of overall performance.

As can be seen from Table 2, [27] and [16] show best performance on MSRA and NYU respectively. Compared to [27], our method achieves a very close result on MSRA dataset and a much better result on NYU dataset. The main reason is that NYU dataset is captured by structure light camera which shows many defective pixels (black holes) on the depth image. These black holes significantly influence the convolutional operation in 2D space because of large responsive value at edge, while 3D CNN is tolerative to these defective pixels due to the binary data representation in voxel space. Meanwhile, our method outperforms another 3D CNN-based V2V-Net [16] on MSRA dataset due to the skeleton constraint which is proved

effective in section 4.3. Additionally, the running speed of our system during evaluation is about 10.8 fps compared to 3.5 fps in [16] with the same GPU type. The qualitative results are shown in Fig 2.

5   Conclusion

In this paper, we propose a structure-aware 3D hourglass network to estimate hand pose from single depth image. To fully utilize the 2.5D depth data, we re-project it into 3D space and regress to 3D joint heatmap by 3D hourglass network. Despite joint loss, bone loss is introduced into the network as intermediate supervision to explicitly model skeleton constraints. Experimental results show that 3D data representation, 3D heatmap and skeleton constraint all contribute a lot to the final result. And our result on two datasets is on par with state-of-the-art result.

References

[1] Tzimiropoulos G Bulat, A. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732, 2016.

[2] Y. E. Kara C. Keskin, F. KÄśracÂÿ and L. Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *In European Conference on Computer Vision*, pages 852–863, 2012.

[3] A. Yao C. Wan and L. Van Gool. Hand pose estimation from local surface normals. In *In European Conference on Computer Vision*, pages 554–569, 2016.

[4] Xinghao Chen, Guijin Wang, Hengkai Guo, and Cairong Zhang. Pose guided structured region ensemble network for cascaded hand pose estimation. *arXiv preprint arXiv:1708.03416*, 2017.

[5] Yang W. Ouyang W. Ma C. Yuille-A.L. Wang X Chu, X. Multicontext attention for human pose estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5669–5678, 2017.

[6] A. Tejani D. Tang, H. Jin Chang and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 3786–3793, 2014.

[7] P. Kohli C. Keskin T.-K. Kim D. Tang, J. Taylor and J. Shotton. Opening the black box: Hierarchical sampling optimization for estimating human hand pose. In *In IEEE International Conference on Computer Vision*, pages 3325–3333, 2015.

[8] T.-H. Yu D. Tang and T.-K. Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *In IEEE International Conference on Computer Vision*, pages 3224–3231, 2013.

[9] B. Andres M. Andriluka E. Insafutdinov, L. Pishchulin and B. Schiele. Deepercut: A deeper, stronger, and faster multiperson pose estimation model. In *European Conference on Computer Vision*, pages 34–50, 2016.

[10] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3593–3601, 2016.

[11] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 5, 2017.

[12] Hengkai Guo, Guijin Wang, Xinghao Chen, Cairong Zhang, Fei Qiao, and Huazhong Yang. Region ensemble network: Improving convolutional network for hand pose estimation. *arXiv preprint arXiv:1702.02447*, 2017.

[13] J. Yuan H. Liang and D. Thalmann. Parsing the hand in depth images. pages 1241–1253, 2014.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] S. Tang B. Andres M. Andriluka P. V. Gehler L. Pishchulin, E. Insafutdinov and B. Schiele. Deepcut: joint subset partition and labeling for multi person pose estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4937, 2016.

[16] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *CVPR*, volume 2, 2018.

[17] Natalia Neverova, Christian Wolf, Florian Nebout, and Graham W Taylor. Hand pose estimation through semi-supervised and weakly-supervised learning. *Computer Vision and Image Understanding*, 164:56–67, 2017.

[18] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[19] Markus Oberweger and Vincent Lepetit. Deepprior++: Improving fast and accurate 3d hand pose estimation. In *ICCV workshop*, volume 840, page 2, 2017.

[20] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015.

[21] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3633–3642. ACM, 2015.

[22] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 824–832, 2015.

[23] Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust articulated-icp for real-time hand tracking. In *Computer Graphics Forum*, volume 34, pages 101–114. Wiley Online Library, 2015.

[24] Arjun; LeCun Yann; Bregler Christoph Tompson, Jonathan; Jain. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2014.

[25] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):169, 2014.

[26] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. 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)*. IEEE, 2017.

[27] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5147–5156, 2018.

[28] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016.

[29] H.Li W.Yang, W.Ouyang and X.Wang. End-to-end jearning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3073–3082, 2016.