SRN: Stacked Regression Network for Real-time 3D Hand Pose Estimation Supplementary Material

Pengfei Ren
rpf@bupt.edu.cn
Haifeng Sun
sunhaifeng_1@ebupt.com
Qi Qi
qiqi@ebupt.com
Jingyu Wang
wangjingyu@bupt.edu.cn
Weiting Huang
huangweiting@ebupt.com

State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications,
Beijing, China

1 Qualitative Results

Some qualitative results for NYU [6], ICVL [5] and MSRA [4] datasets are shown in Fig. 1. specifically, in the last three columns, we show some error annotations in ICVL [5] and MSRA [4] datasets. On Hands17 [8], we chose some special cases, including low-quality images, extreme views and self-occlusion, to demonstrate the robustness of our method. As is shown in Fig. 2, for the complex cases, our method can still obtain accurate and reasonable pose, which indicates that our method can capture global constraints and correlations among different joint well.

2 Impact of Refinement

We use a 3 stacked network to further evaluate the impact of the refining stage. Fig. 3 and Fig. 4 shows some examples of the iterative process on Hands17 [8] dataset with inaccurate initialization and partial error initialization, respectively. The first row shows the results of the initialize hand pose, the second to third rows show the refined results on stage $1-2$. As is shown in Fig. 3, for some pose and view, initial regression results are inaccurate, which is mainly manifested by the fact that some joints fall outside the hand area. In this case, the inaccurate joints will be fine-tuned in subsequent stages. Furthermore, for some initial estimates with obvious errors in Fig. 4, our method can still obtain satisfying results by re-predicting the joints that are obviously unreasonable.

Method	Ours	V2V-PoseNet	DenseReg	Point-to-Point
GPU	2080 Ti	Titan X	Titan X	Titan Xp
Runtime	3.8 ms	285.7 ms	36 ms	23.9 ms
FPS	263.1	3.5	27.8	41.8

Table 1: Comparison of runtime with state-of-the-art methods

3 Additional Comparison on Runtime

During the testing stage, similar to the [1, 2, 3], we did not take into account the time of intercepting the hand area. Our 2-stack model takes 3.8 ms for one frame in average (263.1 FPS) on a single GeForce RTX 2080 Ti GPU with a batch size 1. Table 1 shows a comparison of runtime to the state-of-the-art methods [1, 2, 3, 7]. Our method outperforms all previous state-of-the-art approaches and the inference time is much less than these methods.

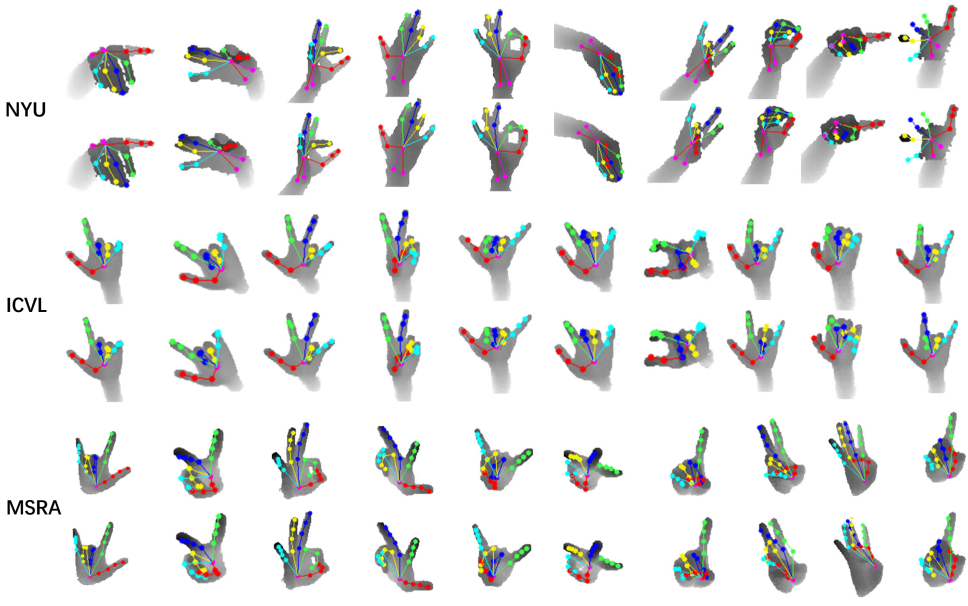

Figure 1: Qualitative results for NYU [6], ICVL [5] and MSRA [4] datasets. We show hand joint locations on depth images. Different hand joints and bones are visualized with different colors. The ground truth hand joint locations are presented in the second row.

References

[1] Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Liuhao Ge, Zhou Ren, and Junsong Yuan. Point-to-point regression pointnet for 3d hand

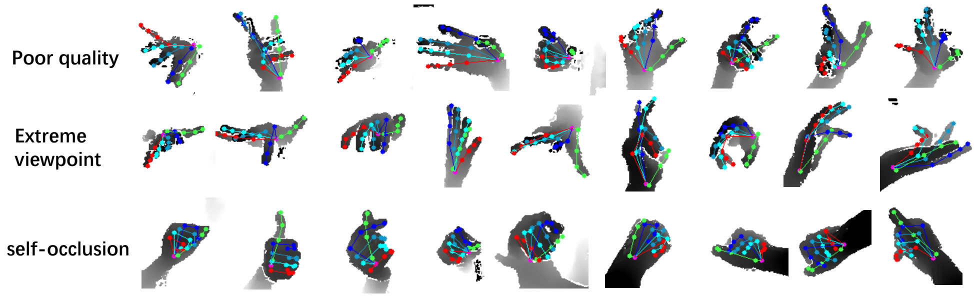

Figure 2: Qualitative results on Hands17 [8] for complex cases.

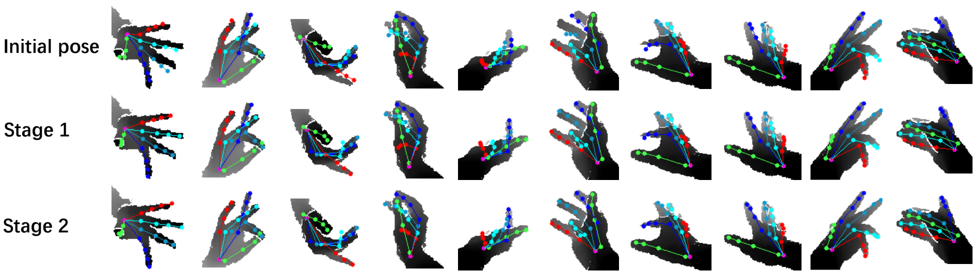

Figure 3: Qualitative results on Hands17 [8] of different stages with inaccurate initialization.

pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[3] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[5] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[6] 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.

[7] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

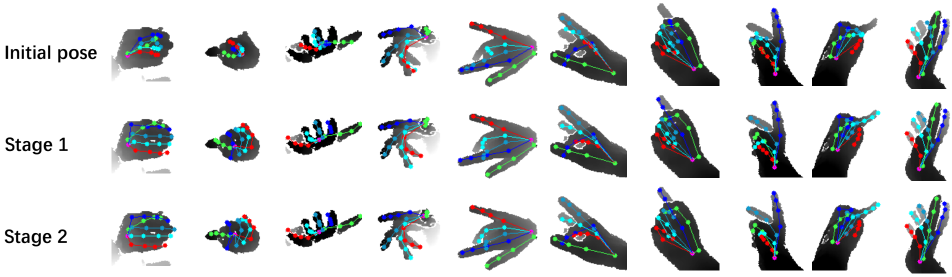

Figure 4: Qualitative results on Hands17 [8] of different stages with error initialization.

[8] Shanxin Yuan, Qi Ye, Guillermo Garcia-Hernando, and Tae-Kyun Kim. The 2017 hands in the million challenge on 3d hand pose estimation. *arXiv preprint arXiv:1707.02237*, 2017.