Supplementary Material of "V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map"

Gyeongsik Moon
ASRI, Seoul National University
mks0601@snu.ac.kr

Ju Yong Chang
Kwangwoon University
juyong.chang@gmail.com

Kyoung Mu Lee
ASRI, Seoul National University
kyoungmu@snu.ac.kr

In this supplementary material, we present more experimental results that could not be included in the main manuscript due to the lack of space.

1. Quantitative comparison with state-of-the-art methods

We compared the performance of the V2V-PoseNet on the three 3D hand pose estimation datasets (ICVL [18], NYU [19], and MSRA [17]) with most of the state-of-the-art methods, which include latent random forest (LRF) [18], cascaded hand pose regression (Cascade) [17], DeepPrior with refinement (DeepPrior) [14], feedback loop training method (Feedback) [15], hand model based method (DeepModel) [25], multi-view CNN (MultiView) [5], DISCO [1], Hand3D [4], lie-x group based method (Lie-X) [21], improved DeepPrior (DeepPrior++) [13], region ensemble network (REN-4$\times$6$\times$6 [8], REN-9$\times$6$\times$6 [7]), CrossingNets [20], pose-guided REN (Pose-REN) [3], global-to-local prediction method (Global-to-Local) [11], classification-guided approach (Cls-Guide) [22], 3DCNN based method (3DCNN) [6], occlusion aware based method (Occlusion) [12], and CDO [10]. Some reported results of previous works [3, 7, 8, 13–15, 18, 21, 25] are calculated by prediction labels available online. Other results [1, 4–6, 10–12, 17, 20, 22] are calculated from the tables of their papers.

As shown in Table 1, our method outperforms all existing methods on the three 3D hand pose estimation datasets.

We also compared the performnace of the V2V-PoseNet on the 3D human pose estimation dataset (ITOP [9]) with more methods than used in the main manuscript. The comparison includes random forest-based method (RF [16]), random tree walk (RTW [24]), IEF [2], viewpoint-invariant feature-based method (VI) [9], and REN-9x6x6 [7]. The scores of the methods are obtained from [7, 9].

As shown in Table 2, our V2V-PoseNet outperforms all the existing methods in front- and top-view.

2. Qualitative results

We report some qualitative results of the V2V-PoseNet on the four 3D hand pose estimation datasets (ICVL [18], NYU [19], MSRA [17], and HANDS 2017 frame-based 3D hand pose estimation challenge dataset [23]) and one 3D human pose estimation dataset (ITOP- Front and Top Views [9]). The results are shown in Figures 1, 2, 3, 4, 5, and 6, respectively.

We also attach videos generated from the test set of the ICVL, NYU, and MSRA datasets.

Methods	Mean error (mm)
LRF	12.58
DeepModel	11.56
Hand3D	10.9
CDO	10.5
DeepPrior	10.4
CrossingNets	10.2
Cascade	9.9
JTSC	9.16
DeepPrior++	8.1
REN-4x6x6	7.63
REN-9x6x6	7.31
Pose-REN	6.79
V2V-PoseNet (Ours)	**6.28**

(a) ICVL

Methods	Mean error (mm)
DISCO	20.7
DeepPrior	19.73
Hand3D	17.6
DeepModel	17.04
JTSC	16.8
Feedback	15.97
Global-to-Local	15.60
Lie-X	14.51
3DCNN	14.1
REN-4x6x6	13.39
REN-9x6x6	12.69
DeepPrior++	12.24
Pose-REN	11.81
V2V-PoseNet (Ours)	**8.42**

(b) NYU

Methods	Mean error (mm)
Cascade	15.2
Cls-Guide	13.7
MultiView	13.2
Occlusion	12.8
CrossingNets	12.2
REN-9x6x6	9.7
DeepPrior++	9.5
Pose-REN	8.65
V2V-PoseNet (Ours)	**7.59**

(c) MSRA

Table 1: Comparison of the proposed method (V2V-PoseNet) with state-of-the-art methods on the three 3D hand pose datasets. Mean error indicates the average 3D distance error.

	mAP (front-view)						mAP (top-view)					
Body part	RF	RTW	IEF	VI	REN-9x6x6	V2V-PoseNet (Ours)	RF	RTW	IEF	VI	REN-9x6x6	V2V-PoseNet (Ours)
Head	63.8	97.8	96.2	98.1	**98.7**	98.29	95.4	**98.4**	83.8	98.1	98.2	**98.4**
Neck	86.4	95.8	85.2	97.5	**99.4**	99.07	98.5	82.2	50.0	97.6	98.9	**98.91**
Shoulders	83.3	94.1	77.2	96.5	96.1	**97.18**	89.0	91.8	67.3	96.1	96.6	**96.87**
Elbows	73.2	77.9	45.4	73.3	74.7	**80.42**	57.4	80.1	40.2	**86.2**	74.4	79.16
Hands	51.3	**70.5**	30.9	68.7	55.2	67.26	49.1	76.9	39.0	**85.5**	50.7	62.44
Torso	65.0	93.8	84.7	85.6	98.7	**98.73**	80.5	68.2	30.5	72.9	**98.1**	97.78
Hip	50.8	80.3	83.5	72.0	91.8	**93.23**	20.0	55.7	38.9	61.2	85.5	**86.91**
Knees	65.7	68.8	81.8	69.0	89.0	**91.80**	2.6	53.9	54.0	51.6	70.0	**83.28**
Feet	61.3	68.4	80.9	60.8	81.1	**87.6**	0.0	28.7	62.4	51.5	41.6	**69.62**
Mean	65.8	80.5	71.0	77.4	84.9	**88.74**	47.4	68.2	51.2	75.5	75.5	**83.44**

Table 2: Comparison of the proposed method (V2V-PoseNet) with state-of-the-art methods on the front and top views of the ITOP dataset. mAP represents the mean average precision.

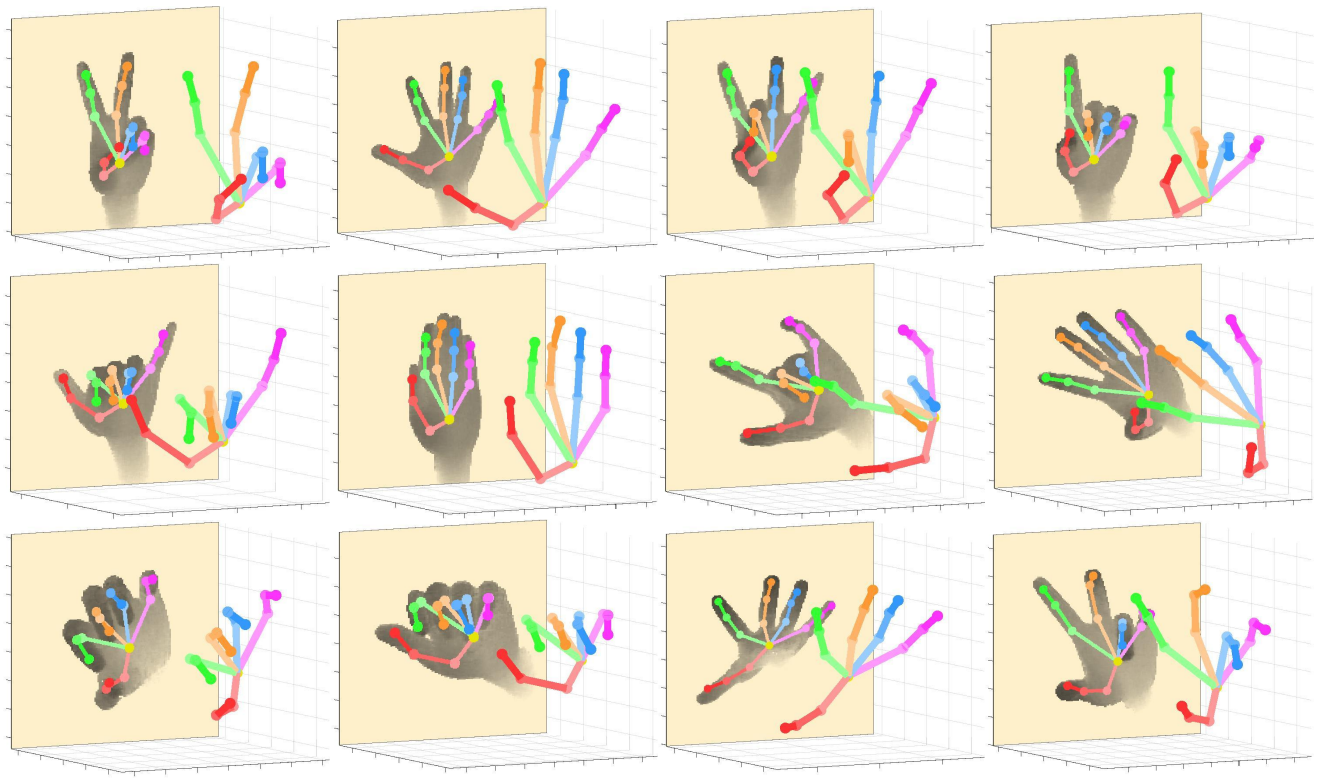

Figure 1: Qualitative results of our V2V-PoseNet on the ICVL dataset. Backgrounds are removed to make them visually pleasing.

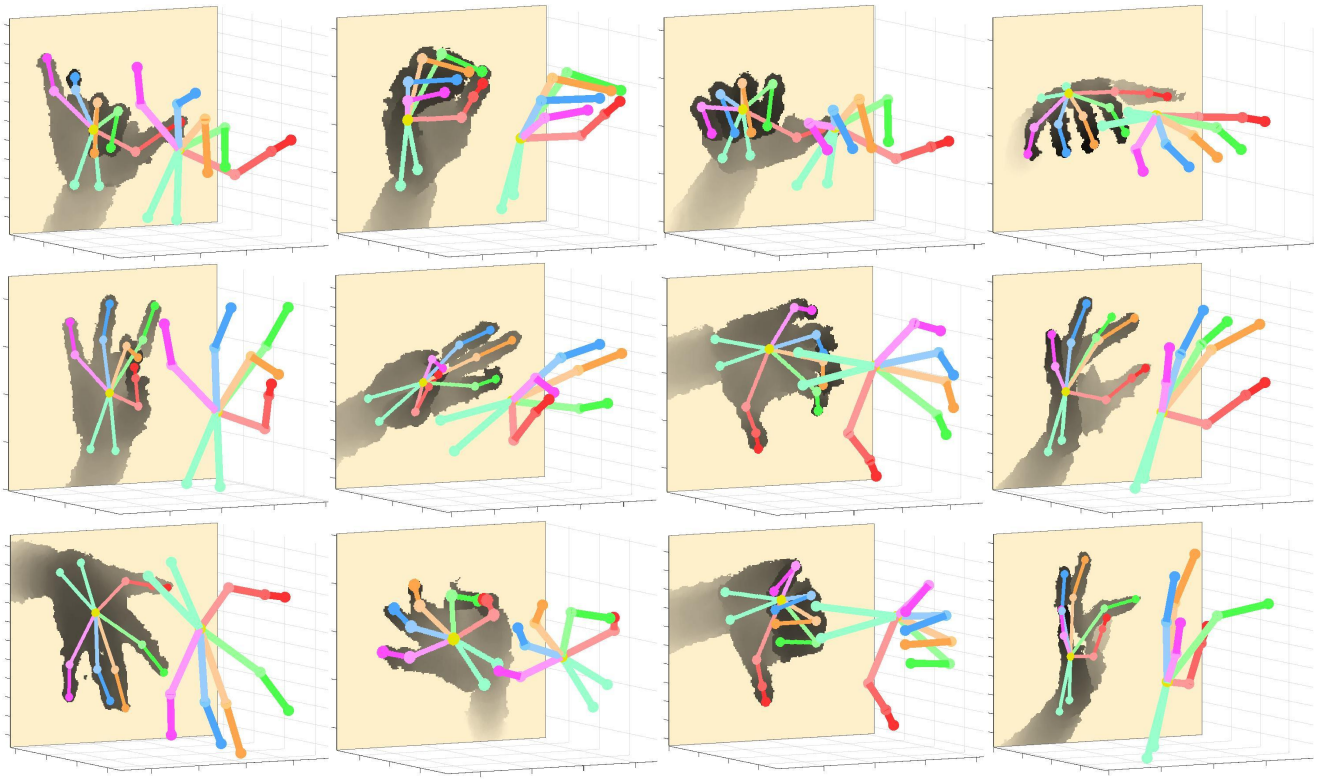

Figure 2: Qualitative results of our V2V-PoseNet on the NYU dataset. Backgrounds are removed to make them visually pleasing.

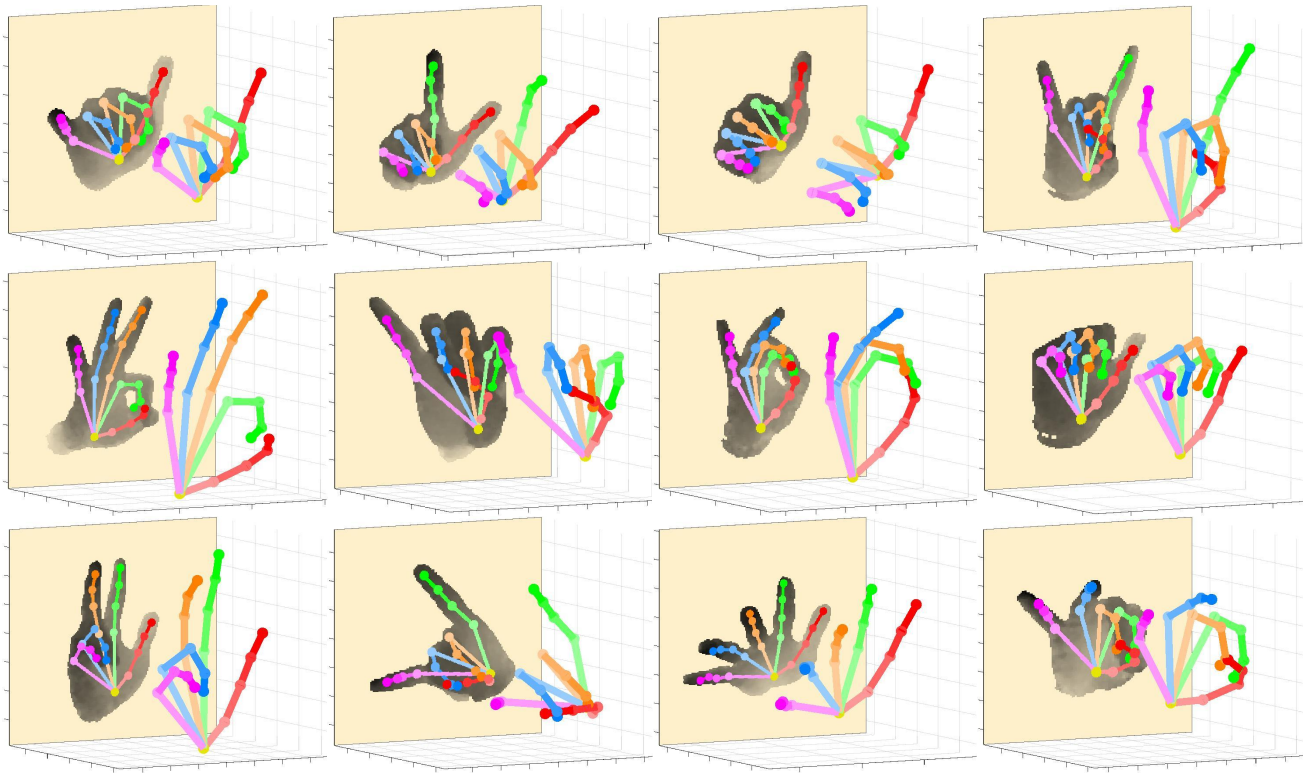

Figure 3: Qualitative results of our V2V-PoseNet on the MSRA dataset. Backgrounds are removed to make them visually pleasing.

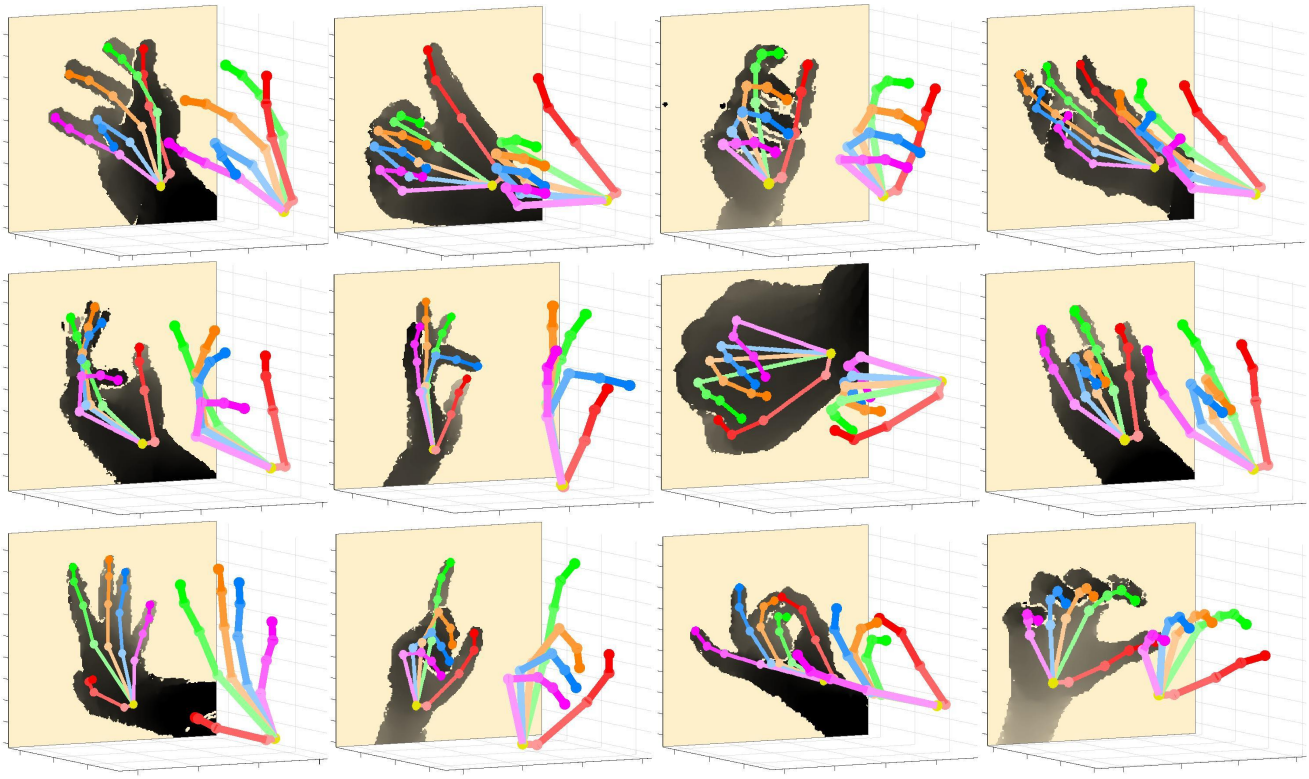

Figure 4: Qualitative results of our V2V-PoseNet on the HANDS 2017 frame-based 3D hand pose estimation challenge dataset. Backgrounds are removed to make them visually pleasing.

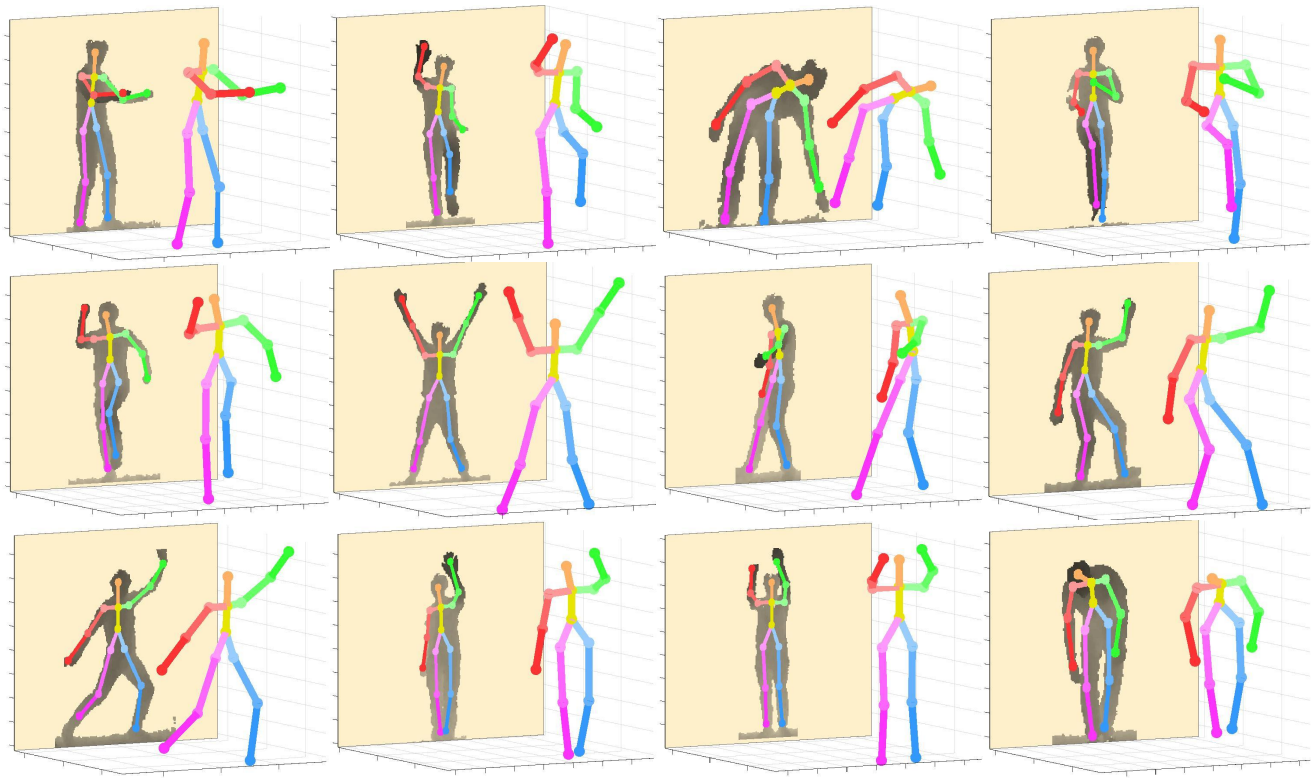

Figure 5: Qualitative results of our V2V-PoseNet on the ITOP dataset (front-view). Backgrounds are removed to make them visually pleasing.

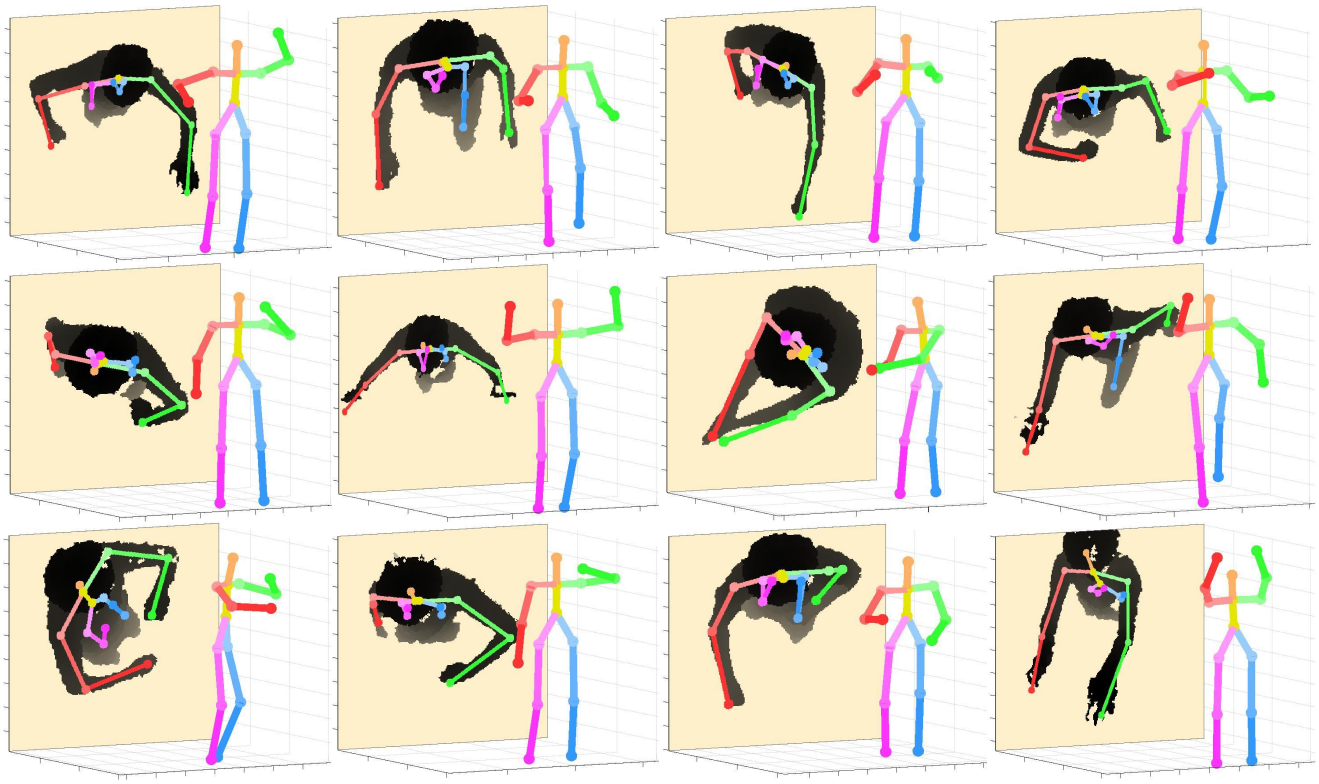

Figure 6: Qualitative results of our V2V-PoseNet on the ITOP dataset (top-view). Backgrounds are removed to make them visually pleasing.

References

[1] D. Bouchacourt, P. K. Mudigonda, and S. Nowozin. Disco nets: Dissimilarity coefficients networks. In *Advances in Neural Information Processing Systems*, pages 352–360, 2016.

[2] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4733–4742, 2016.

[3] X. Chen, G. Wang, H. Guo, and C. Zhang. Pose guided structured region ensemble network for cascaded hand pose estimation. *arXiv preprint arXiv:1708.03416*, 2017.

[4] X. Deng, S. Yang, Y. Zhang, P. Tan, L. Chang, and H. Wang. Hand3d: Hand pose estimation using 3d neural network. *arXiv preprint arXiv:1704.02224*, 2017.

[5] L. Ge, H. Liang, J. Yuan, and D. Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3593–3601, 2016.

[6] L. Ge, H. Liang, J. Yuan, and D. Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[7] H. Guo, G. Wang, X. Chen, and C. Zhang. Towards good practices for deep 3d hand pose estimation. *arXiv preprint arXiv:1707.07248*, 2017.

[8] H. Guo, G. Wang, X. Chen, C. Zhang, F. Qiao, and H. Yand. Region ensemble network: Improving convolutional network for hand pose estimation. *IEEE International Conference on Image Processing*, 2017.

[9] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In *European Conference on Computer Vision*, pages 160–177. Springer, 2016.

[10] P. Krejov, A. Gilbert, and R. Bowden. Guided optimisation through classification and regression for hand pose estimation. *Computer Vision and Image Understanding*, 155:124–138, 2017.

[11] M. Madadi, S. Escalera, X. Baro, and J. Gonzalez. End-to-end global to local cnn learning for hand pose recovery in depth data. *arXiv preprint arXiv:1705.09606*, 2017.

[12] M. Madadi, S. Escalera, A. Carruesco, C. Andujar, X. Baró, and J. Gonzàlez. Occlusion aware hand pose recovery from sequences of depth images. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 230–237. IEEE, 2017.

[13] M. Oberweger and V. Lepetit. Deepprior++: Improving fast and accurate 3d hand pose estimation. In *IEEE International Conference on Computer Vision*, Oct 2017.

[14] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. *Computer Vision Winter Workshop*, pages 21–30, 2015.

[15] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *IEEE International Conference on Computer Vision*, pages 3316–3324, 2015.

[16] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[17] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 824–832, 2015.

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

[19] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 33(5):169, 2014.

[20] C. Wan, T. Probst, L. Van Gool, and A. Yao. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[21] C. Xu, L. N. Govindarajan, Y. Zhang, and L. Cheng. Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups. *International Journal of Computer Vision*, pages 1–25, 2017.

[22] H. Yang and J. Zhang. Hand pose regression via a classification-guided approach. In *Asian Conference on Computer Vision*, pages 452–466. Springer, 2016.

[23] S. Yuan, Q. Ye, G. Garcia-Hernando, and T.-K. Kim. The 2017 hands in the million challenge on 3d hand pose estimation. *arXiv preprint arXiv:1707.02237*, 2017.

[24] H. Yub Jung, S. Lee, Y. Seok Heo, and I. Dong Yun. Random tree walk toward instantaneous 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2467–2474, 2015.

[25] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei. Model-based deep hand pose estimation. *International Joint Conference on Artificial Intelligence*, pages 2421–2427, 2016.