3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images Supplementary Material

Liuhao Ge[1], Hui Liang[1,2,*], Junsong Yuan[1,†], Daniel Thalmann[1]
[1]Institute for Media Innovation, Interdisciplinary Graduate School, Nanyang Technological University
[2]Institute of High Performance Computing, A*STAR, Singapore
ge0001ao@ntu.edu.sg, lianghui@ihpc.a-star.edu.sg, {jsyuan, danielthalmann}@ntu.edu.sg

1. Network Architectures for Self-comparison

We present all the 3D CNN models in the experiment section. We experiment with projective D-TSDF volumes with different resolution values: 16, 32 and 64. Figure 1a presents the network architecture when the input is projective D-TSDF volumes with 32×32×32 resolution. We use this 3D CNN model to compare with state-of-the-art methods on the MSRA dataset [2] and the NYU dataset [3]. However, when the volume resolution is 16×16×16 or 64×64×64, the network architecture is different with that in Figure 1a. Figure 1b presents the network architecture when the input is projective D-TSDF volumes with 16×16×16 resolution. We reduce a convolutional layer and a max pooling layer in this network. Figure 1c presents the network architecture when the input is projective D-TSDF volumes with 64×64×64 resolution. We add a convolutional layer and a max pooling layer in this network.

We also experiment with different TSDF types: accurate TSDF, projective TSDF and projective D-TSDF. When the input volume is accurate/projective TSDF which has only one channel, the parameters of the network architecture in Figure 1a should be modified to adapt to the input with one channel. Figure 1d presents the network architecture when the input is accurate/projective TSDF volumes with 32×32×32 resolution. Since the number of input channel is 1 instead of 3, we divide the numbers of output channels for the convolutional layers by 3.

2. Qualitative Results

Figure 2 presents some qualitative results for hand pose estimation on the MSRA dataset [2] and the NYU dataset [3]. We re-implement the multi-view CNN based hand pose estimation method [1] on both two datasets. As can be seen, the estimation results of our 3D CNN based method are overall better than the results of multi-view CNN based method.

3. Additional Experiments

We conduct additional self-comparison experiments to further validate the effectiveness of our proposed method. We first experiment with the occupancy grid. We conduct this experiment on the whole MSRA dataset [2]. We set the value of occupied voxel as 1, and that of unoccupied voxel as 0. As shown in Figure 3 left, the estimation accuracy of the occupancy grid is worse than that of the projective D-TSDF, which indicates that the projective D-TSDF is more suitable to be used as the volumetric representation. Furthermore, the comparable result achieved by the occupancy grid also shows that our proposed 3D CNN can learn 3D features from different volumetric representations.

Secondly, we experiment with 2D data augmentation. In this experiment, we rotate and stretch the cropped 2D hand depth image. As shown in Figure 3 left, compared with the method without any augmentation, 2D data augmentation can improve the performance evidently and 3D data augmentation can further improve the performance. It is worth noting that there is only a little performance improvement from 2D data augmentation to 3D data augmentation. The reason is that for 2D data augmentation, we only rotate the z axis of the camera's coordinate system, thus the augmented data still follow the distribution of the original data; however, for 3D data augmentation, we not only rotate the z axis, but also rotate x, y axes, thus some 3D points will be occluded from the new viewpoint, and some other 3D points will be missing. Although the augmented point cloud is different with the real point cloud to some extent, the learned CNN can still be adapt to real point clouds and achieve a little better performance than 2D data augmentation.

Finally, we experiment with a deeper network. We train a deeper network having four convolutional layers and four

[*] This work was done when Hui Liang was a research staff at NTU.
[†] Corresponding author

fully-connected layers. Note that in this experiment, we train on subjects P1–P8 with about 68K frames and test on subject P0 with about 8.5K frames in the MSRA dataset [2]. As shown in Figure 3 right, the performance of deeper network is almost the same as the shallower one. Thus, in our implementation, we choose the shallower network which is faster.

References

[1] L. Ge, H. Liang, J. Yuan, and D. Thalmann. Robust 3D hand pose estimation in single depth images: from single-view C-NN to multi-view CNNs. In *CVPR*, 2016.

[2] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *CVPR*, 2015.

[3] 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.

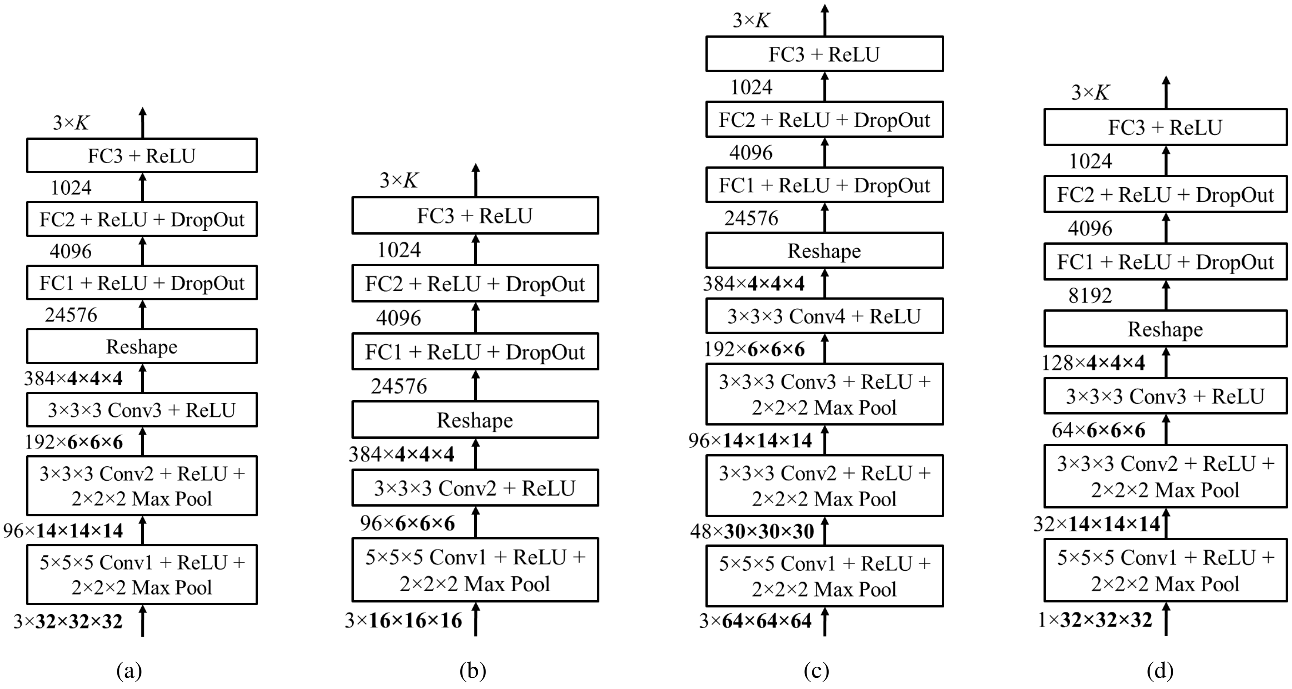

Figure 1: (a) Architecture of 3D CNN with projective D-TSDF, 32×32×32 volume resolution. (b) Architecture of 3D CNN with projective D-TSDF, 16×16×16 volume resolution. (c) Architecture of 3D CNN with projective D-TSDF, 64×64×64 volume resolution. (d) Architecture of 3D CNN with accurate/projective TSDF, 32×32×32 volume resolution.

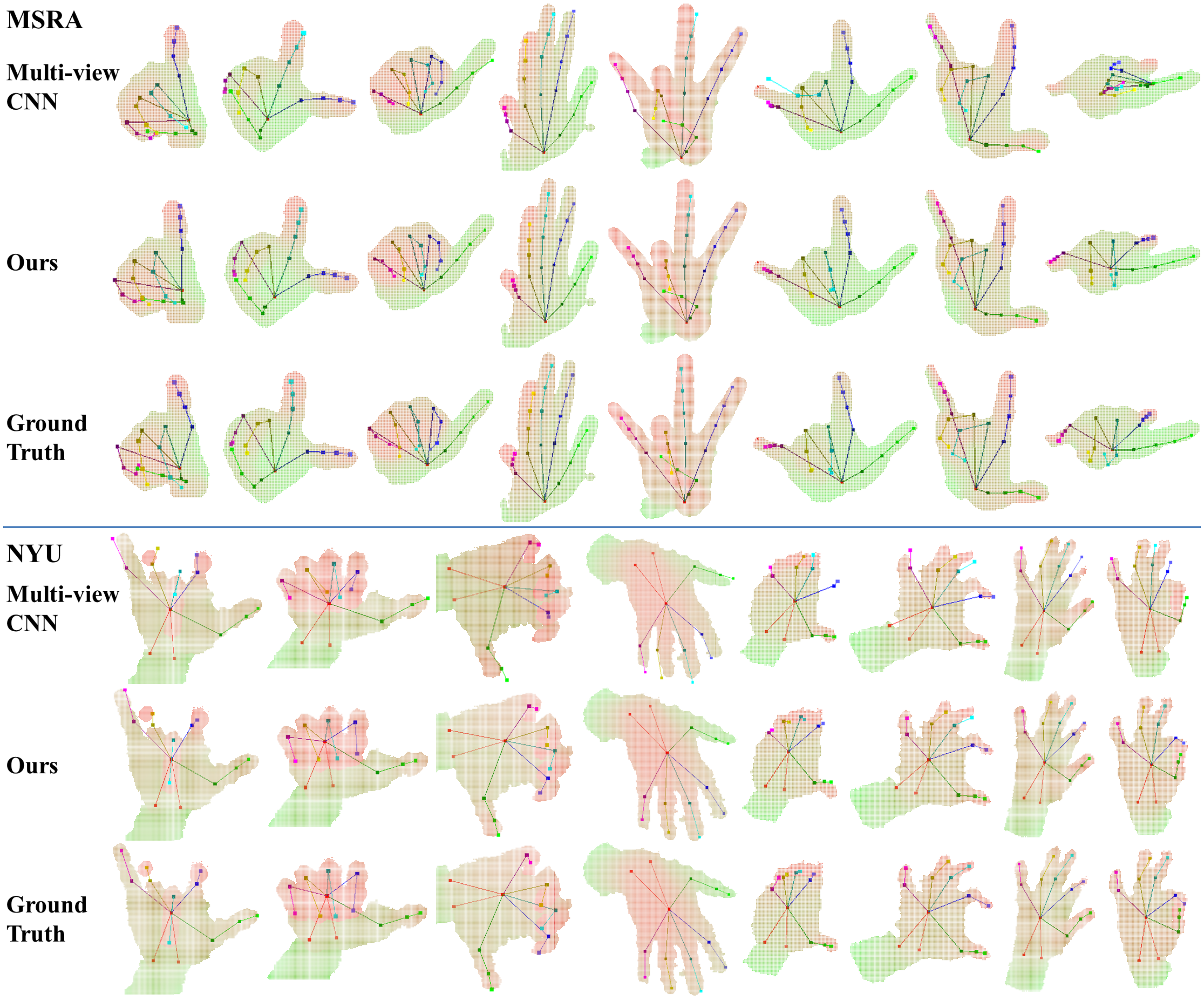

Figure 2: Qualitative results for MSRA dataset [2] and NYU dataset [3]. We compare our 3D CNN based method (in the 2nd line and the 5th line) with the multi-view CNN based method in [1] (in the 1st line and the 4th line). The ground truth hand joint locations are presented in the 3rd line and the 6th line. We show hand joint locations on the depth image. Different hand joints and bones are visualized using different colors. This figure is best viewed in color.

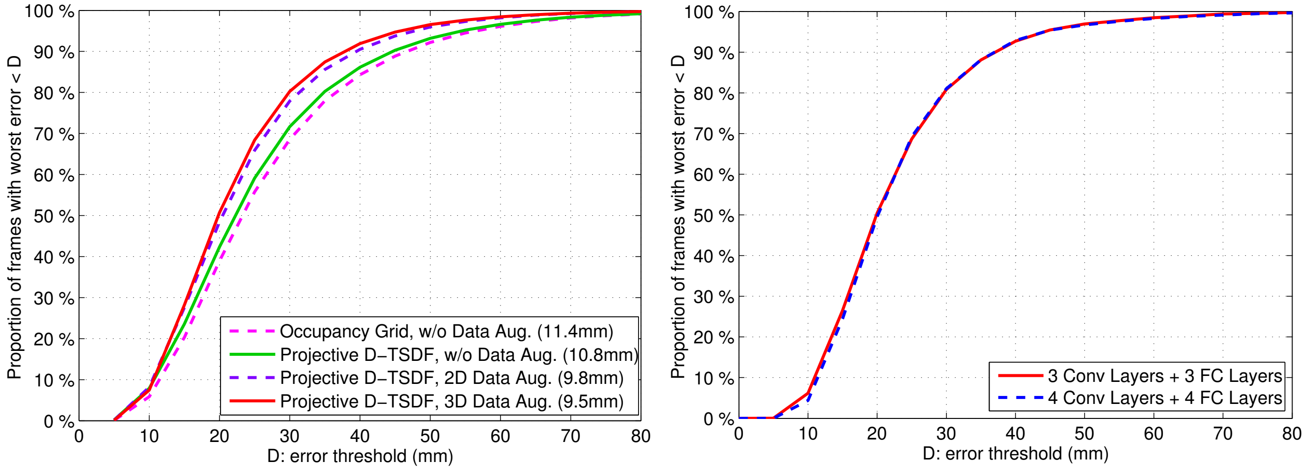

Figure 3: Self-comparison of different methods on MSRA dataset [2]. **Left**: self-comparison with the occupancy grid and the 2D data augmentation. **Right**: self-comparison with a deeper network. This figure is best viewed in color.