Supplementary materials for "JGR-P2O: Joint Graph Reasoning based Pixel-to-Offset Prediction Network for 3D Hand Pose Estimation from a Single Depth Image"

Detailed Flowchart of the Whole Network Architecture

In this paper, we use a two-stage JGR-P2O network for 3D hand pose estimation from a single depth image. The detailed flowchart of the whole network architecture is shown in Figure 1. Given a depth image of a hand, the low-level feature extract module first extracts low-level visual feature. The extracted low-level visual feature goes through an hourglass module, obtaining intermediate local feature representation. Then the GCN-based joint graph reasoning module augments the intermediate local feature representation and obtains the enhanced local feature representation. A 1x1 convolutional layer is performed on the enhanced local feature representation, outputing offset maps for all joints. The low-level visual feature, offset maps, and enhanced local feature representation are combined to form the input to the second hourglass module. We use the outputs of the second stage as the final prediction results.

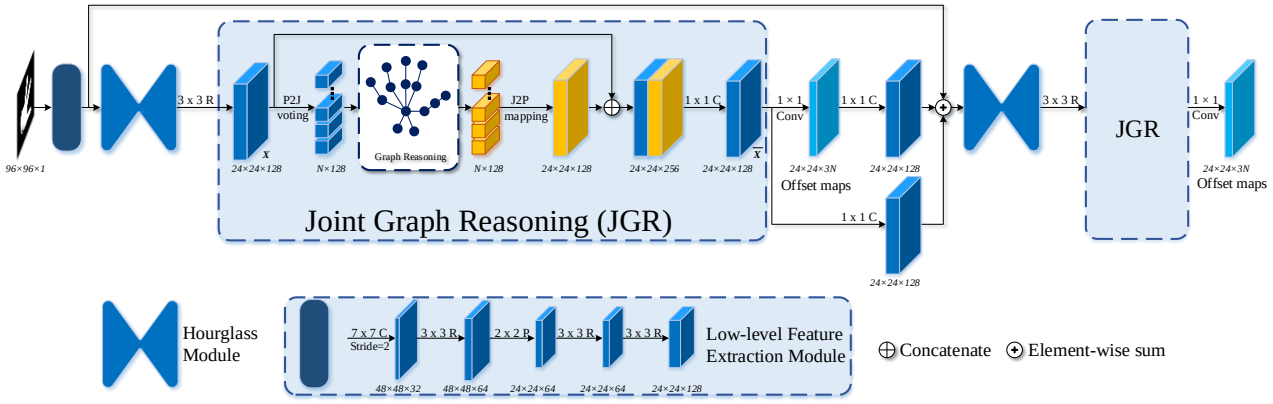

Figure 1: Detailed flowchart of the two-stage JGR-P2O network. The abbreviations Conv, C, P, R indicate convolutional layer without BN and ReLU, convolutional layer with BN and ReLU, pooling layer and residual module respectively.

Per-joint Mean 3D Distance Error

We calculate the mean 3D distance error for each hand joint and compare our method with previous state-of-the-art methods. The results are shown in Figure 2. It can be found that our method can achieve the lowest mean 3D distance errors for 6 joints out of all 16 joints on the ICVL dataset. It can also achieve the lowest mean 3D distance errors for 8 joints out of all 14 joints on the NYU dataset. Furthermore, our method obtains the lowest mean 3D distance error over all joints on both the ICVL and NYU dataset. For the MSRA dataset, our method obtains competitive results compared to previous state-of-the-art dense prediction-based methods, such as DenseReg, V2V-PoseNet, and Point-to-Point.

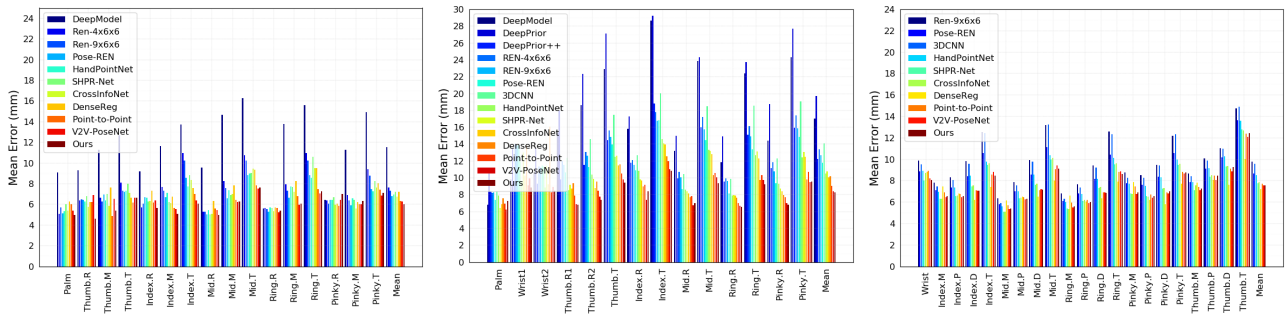

Figure 2: Comparison with previous state-of-the-art methods. The per-joint mean 3D distance errors are presented in this figure. Left: ICVL dataset, Middle: NYU dataset, Right: MSRA dataset.

Qualitative Results

We have made several videos of qualitative results on two datasets: ICVL, NYU. The videos can be found in our attached videos. We would like to point out that the testing images of these two datasets are captured sequentially containing lots of occlusions, motions, and viewpoint changes. Since for both ICVL and NYU the test set contains two sequences, we made a video for each sequence resulting a total of 4 videos. In the videos, ground truth is shown in green, and the estimated pose is shown in red. As shown in these videos, our method can effectively handle the problems of occlusions, motions, and viewpoint changes.

Structure of Skeleton Graph

The structures of skeleton graph on three datasets are shown in Figure 3. There are total 16, 14, and 21 joints in the skeleton graph of the ICVL, NYU and MSRA dataset, respectively.

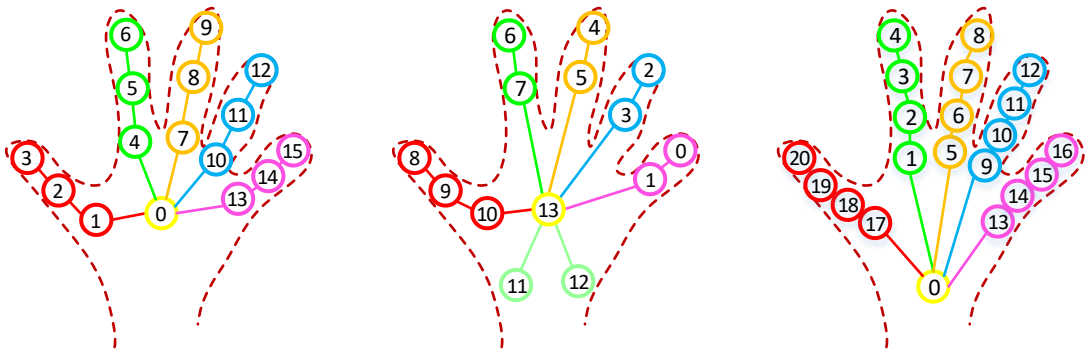

Figure 3: The structures of skeleton graph on different datasets. Left: ICVL dataset, Middle: NYU dataset, Right: MSRA dataset.