Supplementary Material
C3P: Cross-domain Pose Prior Propagation for Weakly Supervised 3D Human Pose Estimation

Cunlin Wu[1], Yang Xiao[1†], Boshen Zhang[2], Mingyang Zhang[1], Zhiguo Cao[1], and Joey Tianyi Zhou[3]

[1] Key Laboratory of Image Processing and Intelligent Control, Ministry of Education, School of AIA, Huazhong University of Science and Technology, China
{cunlin_wu,Yang_Xiao,izmy,zgcao}@hust.edu.cn
[2] YouTu Lab, Tencent
boshenzhang@tencent.com
[3] A*STAR Centre for Frontier AI Reasearch (CFAR)
joey.tianyi.zhou@gmail.com

1 Superiority of our weak supervision approach with 3D projection rays.

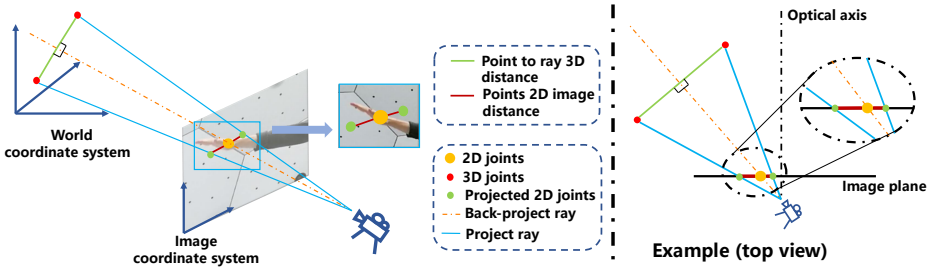

Fig. 1. The left figure shows the phenomenon that, the 2 points of equal distance to the back-projected ray in 3D space are of different distances to ray's native point in 2D image plane. The right figure is the corresponding visualization result from the top view.

In order to further analyze the superiority of our weak supervision signal from RGB image, we show the defect of estimating the inconsistency between 2D and 3D prediction on joint's position in 2D image plane space. As shown in Fig. 1, given a 2D joint within RGB image, its ground-truth (GT) 3D counterpart is on the back-projected ray according to the camera parameters. For two predicted 3D points symmetric to the back-projected ray, no matter where the GT 3D joint is located on the ray, the two predicted points have the same distance to the GT 3D joint in 3D space. Our approach calculates the distance between predicted 3D points and back-projected 3D rays can maintain the error consistency. In

contrast, projecting 3D point to 2D image plane leads to essential inconsistency. As shown in Fig. 1 (right), we show an intuitive example from the top view. The point far from the optical axis is of larger error in image coordinate system, and holds the larger loss in back propagation. This leads to the fact that the loss function cannot reflect the true deviation between the predicted joints and GT joints in 3D space. Inconsistency of the loss function tends to weaken the effectiveness of weak supervision.

2 Per point aggregation detail

In this section, the details of our proposed per point aggregation approach will be illustrated. The difference between our approach and original P2P [2] will also be introduced. For each point in the point cloud, the network outputs the heatmaps and offset fields to predict the location of human joints, following P2P [2]. The heatmap and target offset field are defined as:

$$H(p_i, \phi_j) = \begin{cases} 1 - \|p_i - \phi_j\| / r & \|p_i - \phi_j\| \leq r, \\ 0 & Otherwise, \end{cases} \tag{1}$$

$$U(p_i, \phi_j) = \begin{cases} (p_i - \phi_j) / \|p_i - \phi_j\| & \|p_i - \phi_j\| \leq r, \\ 0 & Otherwise, \end{cases} \tag{2}$$

where p_i is the ith point in point cloud; ϕ_j is the jth joint to be predicted; r is the maximum radius of ball for nearest neighbor search; in our human pose estimation setting, we set r as $80cm/L_{obb}$; L_{obb} is the point cloud normalization parameter according to Handpointnet [1].

Subsequently, the offset vector V from point p_i to predicted joint ϕ_j is defined as:

$$V = r \cdot (1 - H_{ij}) \cdot U_{ij}. \tag{3}$$

The final prediction can be inferred from:

$$\hat{\phi}_j = \frac{\sum_{m=1}^{M} w_m (V_{i_m j} + p_{i_m})}{\sum_{m=1}^{M} w_m}, \tag{4}$$

where i_m is the index of the point corresponding to the m-th largest value of heatmap H_j; We set M as 64 to consider the nearest 64 points in final prediction; w_m is the weight of candidate points according to heatmap value $H_{i_m j}$.

Being different from P2P [2] which defines loss function at heatmaps and unit vector fields, we can define the loss functions according to the joints' 3D locations directly. This makes our method can be trained in an end-to-end way and address the problem that no annotation is provided for generating GT heatmaps and unit vectors in weakly supervised training.

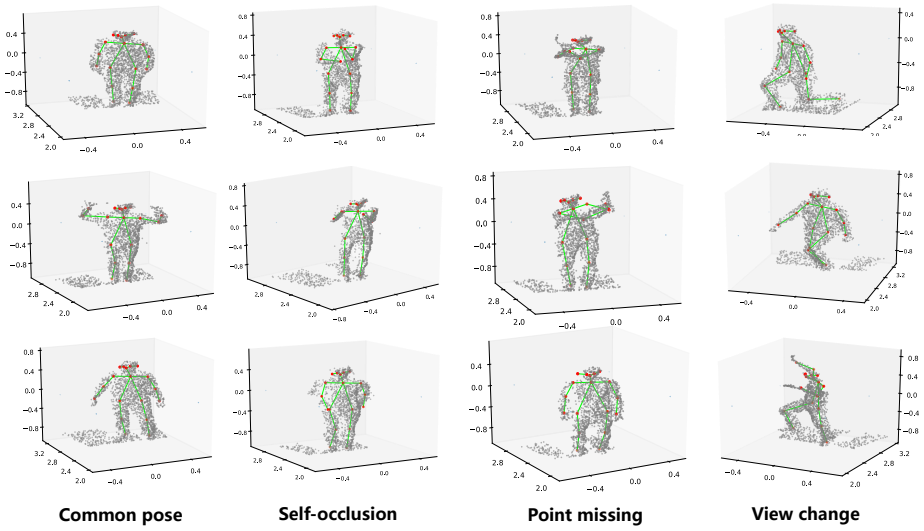

Fig. 2. Qualitative results of C3P on CMU Panoptic Dataset.

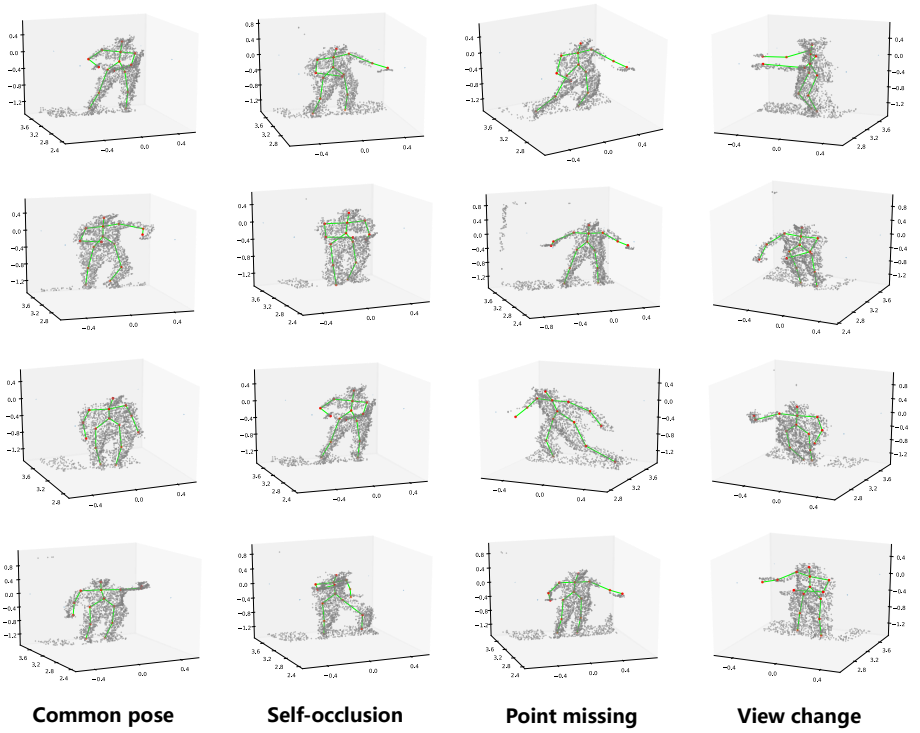

Fig. 3. Qualitative results of C3P on ITOP Dataset.

3 More qualitative results

In this section, as shown in Fig. 2 and Fig. 3, we show more qualitative results of C3P on CMU Panoptic Dataset [4, 5] and ITOP Dateset [3]. Four different situations on 3D human pose estimation in point cloud are shown, including "Common pose", "Self-occlusion", "Point missing" and "View change". It can be observed that, in most cases C3P can acquire acceptable performance.

References

1. Ge, L., Cai, Y., Weng, J., Yuan, J.: Hand pointnet: 3d hand pose estimation using point sets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8417–8426 (2018)
2. Ge, L., Ren, Z., Yuan, J.: Point-to-point regression pointnet for 3d hand pose estimation. In: Proceedings of the European conference on computer vision (ECCV). pp. 475–491 (2018)
3. Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., Fei-Fei, L.: Towards viewpoint invariant 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 160–177 (2016)
4. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
5. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T.S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social interaction capture. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)