A. Regression networks

The network architecture and ground truth preparation for each component are described in detail.

A.1. Backbone network

The backbone network is implemented to include layers 0-2 from ResNet-34 [8] for general image encoding. It outputs a $\frac{w}{8} \times \frac{h}{8} \times 128$ feature map, where h and w indicate height and width of an input image respectively. We choose to maintain $8\times$ downsampling level in the following functional branches to make part-level inference both efficient and capable of handling human parts at a distance.

A.2. Heatmap branch

The heatmap branch is designed to predict $K + 1$ confidence maps corresponding to K body part and the background. The heatmap branch is made of five 3×3 convolutional layers as illustrated in Figure 8. It is worth noting that every convolutional layer mentioned in our method is followed by BN and ReLU layers. To prepare ground-truth part confidence maps $\{H_j^*\}_{j=1}^{K+1}$, we adopt the same method introduced in OpenPose [3], which applies a Gaussian filter at each part location.

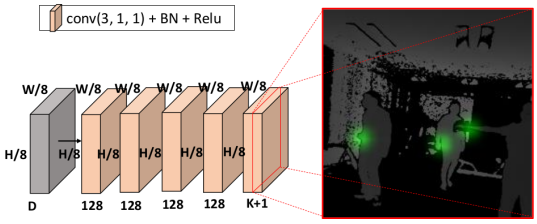

Figure 8. **Heatmap branch.** The heatmap branch predicts K confidence maps for body parts with an additional map for background.

A.3. Depth branch

The depth branch predicts part-wise depth maps, which is meaningful in relieving the effect from raw depth artifacts and in recovering the true depth of a part under occlusion. The network is made of five convolutional layers whose specific architecture is shown in Figure 9.

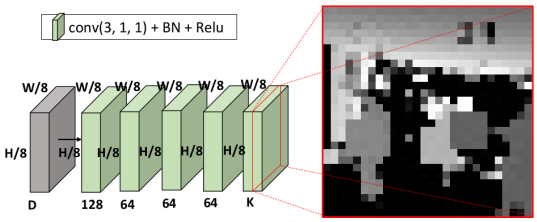

Figure 9. **Depth branch.** The depth branch outputs K depth maps for K body parts, respectively.

To prepare ground-truth depth maps $\{D_j^*\}_{j=1}^{K}$, each map is initialized with the resized raw depth input. The depth values within a 2-pixel-radius disk centered at each part j are overridden with the ground-truth depth of part j, as illustrated in Figure 9. In a multi-person scenario, if a 2D grid position is occupied by masks of more than one part instance, the writing of depth values follows a standard z-buffer rule where the smallest depth value is recorded. In addition, the weight maps $\{W_j^d\}_{j=1}^{K}$ are prepared in the same dimension as the ground-truth depth maps. We use weight 0.9 for a foreground grid while 0.1 for the background.

A.4. TPDF branch

TPDF maps are predicted from the TPDF branch implemented following the architecture as shown in Figure 10. During ground-truth preparation, $\{X_j^*\}_{j=1}^{K}$ and $\{Y_j^*\}_{j=1}^{K}$ are prepared so that the displacement vector at a 2D position points to the closest part position. Specifically, the displacement vector is only non-zero within the truncated range ($r = 2$) from each part position, as shown in Figure 10. The preparation of weight maps $\{W_j^t\}_{j=1}^{K}$ is similar to the process for the depth branch. However, the weights within the truncated mask is set to 1.0 and the rest is set to strict 0.

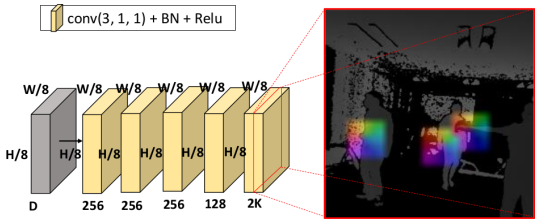

Figure 10. **TPDF branch.** The TPDF branch outputs $2K$ maps of displacement vectors $\{X_j\}_{j=1}^{K}$, $\{Y_j\}_{j=1}^{K}$. The field visualization follows the optical flow standard.

A.5. Global pose network

The global pose network predicts a global pose map from concatenated features from the backbone and functional branches. The network includes four convolutional layers where the first is followed by a max pooling to cast the feature map to $16\times$ downsampling level, as shown in Figure 11.

The ground-truth preparation process is similar to Yolo2 [25]. Specifically, the ground-truth global pose map P^* is prepared so that each grid records five bounding box attributes and a set of pose attributes $\{(dx_j^a, dy_j^a, Z_j^a, v_j^a)\}_{j=1}^{K}$ of the ground-truth pose for each associated anchor a. Specifically, (dx_j^a, dy_j^a) indicates the 2D offsets of part j from the anchor center, Z_j^a indicates the 3D part depth, and v_j^a indicates the visibility of part j. The value of v_j^a is assigned to 1 when the depth from

a global pose part Z_j^a is different from the corresponding depth branch ground-truth in D_j, otherwise it is assigned to 0. The weight map W^p is prepared in the same dimension as P^*. For the dimensions corresponding to bounding box probabilities, 0.9 is applied to the grids associated with ground truth, while 0.1 is assigned to the rest. For the other dimensions, the weights are strictly assigned to 1 or 0. The weight map is designed in such a way because the detection task related to p_b considers both foreground and background while the regression task to other attributes focuses only on foreground.

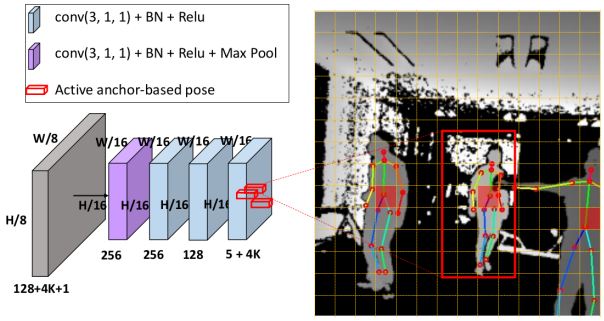

Figure 11. **Global pose network**. The global pose network is composed of four 3×3 convolutional layers, where an additional max-pooling is involved in the first layer. The network outputs an anchor-based global pose map, which is converted to a set of poses after NMS.

B. Depth Augmentation

Given camera intrinsic parameters, the captured depth map, and associated 2D/3D poses, novel depth maps and associated 2D/3D poses can be generated via simulating the camera re-positioned along the principle axis. Specifically, suppose a 3D point (X, Y, Z_0) in the original camera coordinate frame with projection at (x_0, y_0) in the original image is placed at (X, Y, Z_1) in the new camera coordinate frame and be projected to (x_1, y_1) in the new image, we can write the following relationships based on similar triangles:

$$\frac{X}{x_1 - cx} = \frac{Z_1}{f} = \frac{Y}{y_1 - cy} \tag{9}$$

$$\frac{X}{x_0 - cx} = \frac{Z_0}{f} = \frac{Y}{y_0 - cy} \tag{10}$$

where (cx, cy) represents the principle point in both images, and f indicates the focal length. Dividing the first equation by the second, we get:

$$a = \frac{x_0 - cx}{x_1 - cx} = \frac{y_0 - cy}{y_1 - cy} = \frac{Z_1}{Z_0} \tag{11}$$

Thus, a new depth image can be simply generated via randomly sampling a within a reasonable range, and mapping the area defined by the original four image corners to the new locations in the new image. Meanwhile, the depth values of the new image and the associated 2D and 3D body part positions can also be calculated. This depth augmentation method is rather effective. However, it can not simulate the dis-occlusion from a different camera location, such that the augmented depth data can not fully represent the quality of real captured data. In practice, the synthesized depth is directly determined by the original depth and a, whose effect is analysed in Section 4.3.

C. Multi-Person Data Augmentation

Multi-person and background data augmentation plays an important role in training models generalizable to uncontrolled multi-person scenarios. Such augmentation is enabled by the training data of MP-3DHP, which not only includes ground-truth 3D joint positions but also the foreground masks. Specifically, the training set includes human subjects recorded at four different locations relative to the camera plus a set of free-style movements, as shown in Figure 6 (top) and a set of background-only images as shown in the left two images in Figure 6 (bottom). Given a set of background-only images, a human segment from the training set can be used to simply override the pixels within the same region, leading to a background augmented image as shown in Figure 12 (Top). Similarly, human segments from different recording locations can be composed with random background images following the z-buffer rule to generate multi-person augmented images as shown in Figure 12 (Bottom).

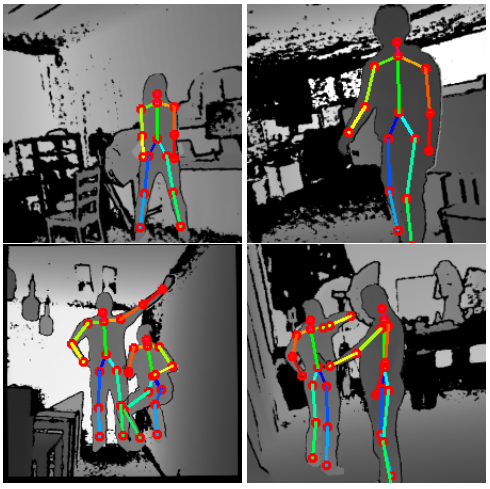

Figure 12. **Augmented training samples.** (Top) Single-person training samples augmented with a random background scene. (Bottom) Augmented multi-person training sample composed from multiple single-person training samples and a random background scene.

There are a few heuristics associated with the simple

augmentation. First, we include no more than two bodies in the multi-person augmentation with an assumption that inter-person occlusion cases between two bodies can well represent the inter-person occlusion cases between more bodies. Second, the straight-forward composition does not consider scene geometry, thus some generated cases appear unrealistic. However, the conflict with scene geometry is not considered a serious issue in training because all the pipelines only adopt convolutional layers learning that only relies on the local context between a body part and the background in its vicinity rather than the whole scene. Finally, there are sensor artifacts around each human segment that can not be perfectly removed. This issue indeed affects the generalization capability of a trained model to the real data. For example, an occluded part from the augmented data is still roughly visible because of the black margin around the human segment, however an occluded part appears truly invisible in real data. Examples of multi-person augmentation and background augmentation are visualized in Figure 12.

D. Effectiveness of data augmentation

The effectiveness of the depth augmentation (**D Aug**) method and the composition augmentation (**C Aug**) are analyzed. The depth augmentation considers different ranges of a. The composition augmentation includes background augmentation (**BG Aug**) and multi-person augmentation (**MP Aug**). Experiments have been conducted with a focus on the multi-person real testing set as these augmentation methods were motivated towards this ultimate task. However, the data augmentation methods are not limited to PoP-Net, but applicable to any method trained on MP-3DHP.

D Aug	C Aug	2D PCK	3D PCK	2D mAP	3D mAP
0.7-1.7	w\o	0.411	0.246	0.374	0.158
0.7-1.7	BG Aug	0.769	0.634	0.748	0.550
0.7-1.7	**MP Aug**	**0.839**	**0.708**	**0.799**	**0.606**
w\o	MP Aug	0.610	0.481	0.617	0.427
0.5-2.5	MP Aug	0.835	0.648	0.785	0.508

Table 7. **Ablation study on data augmentation.**

As observed from Table 7, background augmentation leads to a significant improvement (over 30%) compared to the baseline without any composition augmentation (2nd row vs. 1st row). Multi-person augmentation (3rd row) leads to another leap (about 5%). On the other hand, the specific depth augmentation also plays a critical role in improving the robustness (about 18% increase in 3D mAP, 3rd row vs. 4th row), especially for objects from unobserved scales. However, further extension of the depth augmentation range leads to a drawback in 3D mAP (10%, 5th row vs. 3rd row), which is reasonable because the depth augmentation method can not fully simulate the data far beyond the original captured distance.

E. Detailed running speed analysis

The efficiency of a method is measured in a few different metrics. First, the network complexity is measured in **MACs (G)**, which directly relates to the network inference time. Second, a method's average running time on an image including a single person (**SP**) is reported in milliseconds per image (*ms/im*). This metric considers not only network inference time, but also the essential pre-process to provide bounding boxes or the post-process to extract human poses. Third, a method's average running time on images including multiple people (**MP**) is also reported in milliseconds per image (*ms/im*). Finally, a method's average running speed on images including multiple people is measured by *fps* which is equivalent to the metric in *ms* per image on MP data. Every method has been tested on a single RTX 2080Ti GPU, and is evaluated in all the metrics as shown in Table 8.

	Yolo-Pose+	Open-Pose+	A2J	PoP-Net
MACs(G)	4.4	6.7	16.6	6.2
SP (ms/im)	4.5	20	14	11
MP (ms/im)	4.5	21	32	11
MP (fps)	223	48	32	91

Table 8. **Running time analysis on multi-person data.**

As observed from MACs (G) scores, Yolo-Pose+ has the lightest network, while A2J has significantly more complex network compared with others. However, consider the pipeline running time, Open-Pose+ is much slower than the others on single-person images. This indicates that the part association post-process involved in Open-Pose+ is a much heavier process compared with the simple post-process used in Pop-Net. On the other hand, although A2J uses a more complex network, it almost has no post-processing cost so that its efficiency on single-person images is even better than Open-Pose+. Finally, as observed from multi-person pipeline running time and speed, the efficiency of A2J drops significantly while the other single-shot methods are not affected. Overall, PoP-Net shows significant advantages in efficiency, which almost triples A2J and doubles Open-Pose+ in multi-person scenarios. It can be anticipated that the speed advantage of PoP-Net will be more dominating when more people are involved.

F. Application

For AR/VR applications, we demonstrate that our prediction of 3D human body parts enables the virtual avatar driving where 3D motion capture plays a key role. As shown in Figure 13, we convert a sequence of predicted 3D joint positions into the rotation angles of each joint to drive the animated virtual avatar. The supplemental video shows a frame-by-frame avatar-driving animation, and the result is further smoothed by inter-frame filtering. Here, we show

the result by recovering the rotation angle of each joint, and the pelvis position is fixed in a certain location.

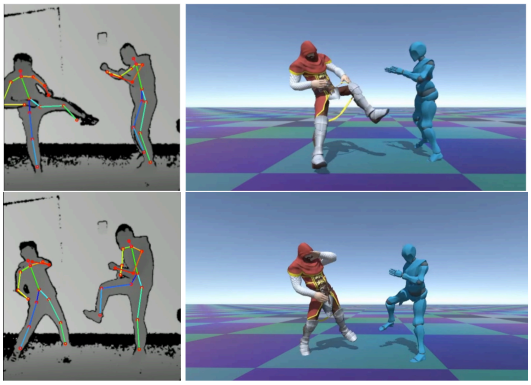

Figure 13. **Virtual avatar driving results.** The left column shows the input depth images and the right column shows the corresponding virtual avatar interaction.

G. Qualitative comparison

In order to demonstrate that PoP-Net achieves the state-of-the-art and the proposed MP-3DHP represents real-world challenges, we compare the predicted poses from competing methods on a set of challenging cases. As shown in Figure 14, we demonstrate visual comparison on: (1) an example including a target human captured far beyond the observation range in the training; (2-3) examples having severe background occlusion; (4-5) examples including multi-person occlusion and considerable truncation; and (6) an example with unobserved poses from training. As observed, PoP-Net in general is more reliable across all these cases. However, all methods failed on some most challenging cases. Such observation indicates that there is still huge room for improvement towards a robust approach in real-world challenges.

H. Visual comparison in videos

To provide a direct visual comparison between different methods, we also provide supplemental material in video format. Specifically, visual results of candidate methods are composed into a video including different multi-person configurations recorded at two different scenes. Within each video frame, Open-Pose+ [3, 14], Yolo-A2J [36] and PoP-Net (ours) are visually compared to Azure Kinect outputs. We chose to compare to Kinect on raw videos for comparison rather than the manually verified testing set in order to show there are cases our method even outperforms Kinect. It is also worth noting that Kinect results are cropped from images with larger FOV compared to the other methods, such that some poses under huge truncation are still accurately visualized.

Figure 14. **Visual comparison of competing methods on challenging cases.**