Supplementary Material

1 Qualitative Study

In Figure 1, we provide a qualitative comparison between our results and those presented by VoxelPose. For this comparison, we chose scenes characterized by pronounced occlusion and obstructions. Our approach leads to reduced erroneous detections and produces more accurate human poses.

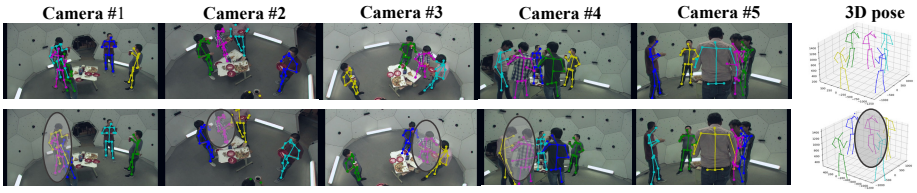

Fig. 1: Qualitative analysis. Both models are evaluated in environments with occlusions and obstructions to compare their performances. Estimated 3D poses and their 2D projections of ours (the 1st row), and Tu et al. (the 2nd row). Wrong prediction poses are highlighted with circles.

2 Additional results on 3D space attention visualization

Additional space attention visualization results are illustrated in Figure 2. We present the results on two different action sequences. The 1st raw represents the space attention scores distribution in the 3D space. The 2nd raw represents the space attention region projected to the 2D image. We divide the feature volume into $10 \times 10 \times 3$ regions, as depicted in the 1st row, the intervals deemed important within the space are concentrated around the locations where people appear. It's crucial to highlight that during the training process, we did not supervise the space attention scores. In the 2nd row, we project the space attention from the space onto a 2D image, where the crucial parts also focus on the areas where people appear. This aligns with our hypothesis, enabling us to use the space attention module to predict the significance of different intervals within the space.

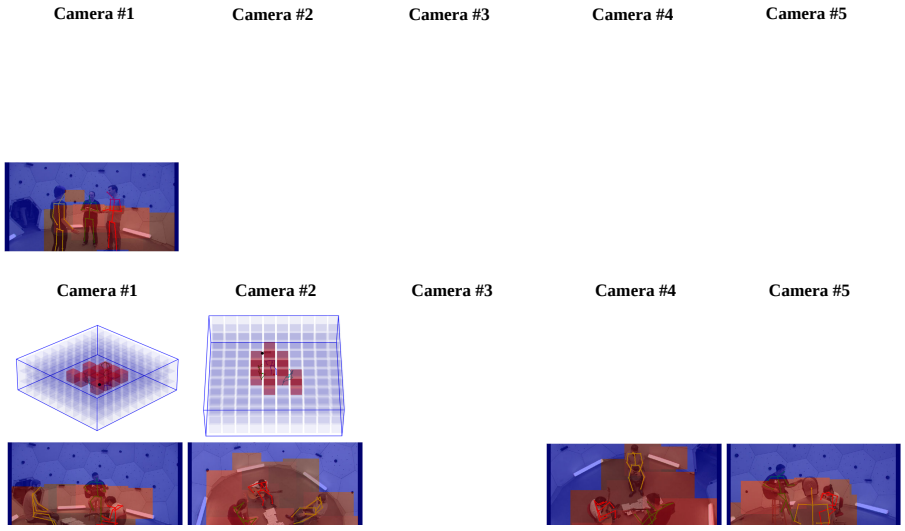

Fig. 2: Additional space attention visualization results on Panoptic datasets.