1 Appendix

Table 1: **Improvements** on MPII validation set.

Method	Input size	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
HRNet-W32 [2]	256 × 192	97.1	95.9	90.3	**86.4**	89.1	**87.1**	**83.3**	90.3
+ PoseTrans (Ours)	256 × 192	**97.2**	**96.2**	**90.9**	86.3	**89.8**	**87.1**	**83.3**	**90.5**
HRNet-W48 [2]	256 × 192	**97.2**	96.1	90.8	86.3	89.3	86.6	83.1	90.4
+ PoseTrans (Ours)	256 × 192	**97.2**	**96.2**	**91.0**	**86.4**	**89.5**	**87.1**	**83.2**	**90.6**

MPII [1] dataset is a popular dataset for evaluating pose estimation models. It contains 40k person samples, each labeled with 16 joints. We followed the standard train/val/test split as in [3] and use PCKh@0.5 for evaluation. The training settings are the same as we use for the MS-COCO dataset. Table 1 shows the performance improvement on the MPII dataset, where HRNet-W32 and HRNet-W48 with the input size of 256 × 192 are adopted as the baselines. From the table, we can observe that PoseTrans consistently boosts the performance of the baselines.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2014) 1
2. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. arXiv preprint arXiv:1902.09212 (2019) 1
3. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. Advances in neural information processing systems **27** (2014) 1