–Supplementary Material– HMOR: Hierarchical Multi-Person Ordinal Relations for Monocular Multi-Person 3D Pose Estimation

1 Effectiveness of Multi-task Learning

To demonstrate the advantage of the multi-task learning strategy, we compare the accuracy of *human detection*, *human-depth estimation*, and *root-relative pose estimation* against the disjoint learning pipeline. Since Moon et al. [5] have reported their results of each module on MuPoTS-3D, we conduct the comparative experiments on this dataset directly. Quantitative results are reported in Table 1, Table 2, Table 3. For *human detection*, we compare with the original Mask R-CNN [1]. For *human-depth estimation*, and *root-relative pose estimation*, we compare with the disjoint learning model [5]. In the training phase, our model is still trained with a multi-task strategy. In the testing phase, we use the same human detection results as Moon et al. [5] (45.0 AP) to extract RoIs for a fair comparison. All models are trained with 400K composite frames from MuCo [4] and additional MSCOCO [2] data.

Table 1. Comparison of human detection

Network	Box AP↑
Mask R-CNN (ResNet-50)	43.8
Ours (ResNet-50)	**47.7**

Table 2. Comparison of human-depth estimation

Network	Depth AP_{25}↑
Moon [5] (ResNet-50)	42.1
Ours (DepthHead)	**48.5**

Table 3. Comparison of root-relative pose estimation

Network	PCK_{rel}↑	AUC_{rel}↑
Moon [5] (ResNet-50)	81.8	39.8
Ours (PoseHead)	**81.9**	**43.3**

2 Detailed results on MuPoTS-3D [4]

Here we report a sequencewise comparison results on MuPoTS-3D [4] dataset in Table 4 and Table 5.

Table 4. Sequence-wise PCK_{abs} comparison with the state-of-the-art method on MuPoTS-3D dataset.

Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg
	Accuracy for all ground truths																				
Moon [5]	**59.5**	**44.7**	51.4	46.0	52.2	27.4	23.7	26.4	**39.1**	23.6	18.3	14.9	38.2	26.5	36.8	23.4	14.4	19.7	18.8	25.1	31.5
Ours	36.8	38.5	**63.7**	**60.4**	**52.3**	**40.2**	**53.6**	**48.3**	36.3	**82.1**	**21.5**	**40.6**	**46.0**	**52.5**	**63.7**	**25.2**	**27.6**	**23.9**	**25.6**	**36.2**	**43.8**
	Accuracy only for matched ground truths																				
Moon [5]	**59.5**	**45.3**	51.4	46.2	53.0	27.4	23.7	26.4	**39.1**	23.6	18.3	14.9	38.2	29.5	36.8	23.6	14.4	20.0	18.8	25.4	31.8
Ours	36.2	39.7	**63.6**	**60.0**	**53.6**	**39.6**	**53.5**	**49.1**	35.7	**81.5**	**21.0**	**40.0**	**48.0**	**59.8**	**63.1**	**24.9**	**27.0**	**24.5**	**25.6**	**44.5**	**44.6**

Table 5. Sequence-wise PCK_{rel} comparison with the state-of-the-art method on MuPoTS-3D dataset.

Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg
	Accuracy for all ground truths																				
Rogez [6]	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1	50.2	51.0	51.6	49.3	56.2	66.5	65.2	62.9	66.1	59.1	53.8
Mehta [4]	81.0	60.9	64.4	63.0	69.1	30.3	65.0	59.6	64.1	83.9	68.0	68.6	62.3	59.2	70.1	80.0	79.6	67.3	66.6	67.2	66.0
Rogez [7]	87.3	61.9	67.9	74.6	78.8	48.9	58.3	59.7	78.1	89.5	69.2	73.8	66.2	56.0	74.1	82.1	78.1	72.6	73.1	61.0	70.6
Moon [5]	94.4	77.5	79.0	81.9	**85.3**	**72.8**	81.9	**75.7**	**90.2**	**90.4**	79.2	79.9	75.1	**72.7**	81.1	89.9	**89.6**	**81.8**	**81.7**	**76.2**	81.8
Ours	**95.5**	**78.0**	**84.8**	**83.9**	84.7	66.2	**83.7**	72.8	86.3	89.0	**81.9**	**83.8**	**78.2**	69.3	**86.7**	**91.8**	88.2	81.5	78.5	72.5	**82.0**
	Accuracy only for matched ground truths																				
Rogez [6]	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
Mehta [4]	81.0	65.3	64.6	63.9	75.0	30.3	65.1	61.1	64.1	83.9	72.4	69.9	71.0	72.9	71.3	83.6	79.6	73.5	78.9	90.9	70.8
Rogez [7]	88.0	73.3	67.9	74.6	81.8	50.1	60.6	60.8	78.2	89.5	70.8	74.4	72.8	64.5	74.2	84.9	85.2	78.4	75.8	74.4	74.0
Moon [5]	94.4	78.6	79.0	82.1	86.6	**72.8**	81.9	**75.8**	**90.2**	**90.4**	79.4	79.9	75.3	**81.0**	81.0	90.7	**89.6**	83.1	**81.7**	77.3	82.5
Ours	**95.5**	**81.8**	**85.5**	**84.2**	**87.8**	66.2	**84.5**	74.9	86.3	89.0	**82.4**	**83.8**	**82.7**	80.0	**86.7**	**92.7**	88.2	**85.5**	80.6	**90.7**	**84.5**

3 Qualitative results

We further provide the qualitative results on 3DPW [3] dataset and more qualitative results of the in-the-wild internet videos in `0148-demo.mp4`. From the failure cases presented in the video, we find out that the predicted depths of the small persons that far away from the camera have obvious jitters. We suspect this is due to the lack of large-scale scenarios in the training data.

References

1. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
2. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
3. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018)
4. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3D pose estimation from monocular rgb. In: 3DV (2018)
5. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3D multi-person pose estimation from a single rgb image. In: ICCV (2019)
6. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: Localization-classification-regression for human pose. In: CVPR (2017)
7. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2D and 3D pose detection in natural images. TPAMI (2019)