Supplementary Material: SMAP: Single-Shot Multi-Person Absolute 3D Pose Estimation

In this supplementary material, we provide more experimental details and results. Additionally, qualitative results on in-the-wild images from the Internet are shown in the supplementary video.

1 More details

1.1 Loss function

There are three output branches of the network, illustrated in Fig. 2. The first branch regresses keypoint heatmaps $\boldsymbol{H}_J$ and PAFs $\boldsymbol{C}$ simultaneously, while the second branch regresses part relative-depth maps $\boldsymbol{H}_{\Delta Z}$. L_2 loss is applied to these two branches. The third branch predicts root depth map $\boldsymbol{H}_{RZ}$. According to 2D location of detected root (x^{root}, y^{root}), we can get the predicted root depth $\boldsymbol{H}_{RZ}(x^{root}, y^{root})$ and compared it with the groundtruth normalized depth $\widetilde{Z}^*$ using L_1 loss. The total loss is computed by weighted summation of all losses. Our loss functions are as follows.

$$
\begin{aligned}
L_{total} = \;& w_{2D} \cdot L_{2D} + w_{\Delta Z} \cdot L_{\Delta Z} + w_{RZ} \cdot L_{RZ} \\
L_{2D} = \;& \sum_{i=1}^{N} \sum_{p} ||\boldsymbol{H}_{J,i}(p) - \boldsymbol{H}^*_{J,i}(p)||_2^2 \quad + \\
& \sum_{i=1}^{2N-2} \sum_{p} ||\boldsymbol{C}_i(p) - \boldsymbol{C}^*_i(p)||_2^2 \\
L_{\Delta Z} = \;& \sum_{i=1}^{N-1} \sum_{p} ||\boldsymbol{H}_{\Delta Z,i}(p) - \boldsymbol{H}^*_{\Delta Z,i}(p)||_2^2 \\
L_{RZ} = \;& \sum_{i=1}^{M} ||\boldsymbol{H}_{RZ}(x_i^{root}, y_i^{root}) - \widetilde{Z}^*_i||_1 \;,
\end{aligned}
$$

where N, M are the number of joints, the number of detected people (root joints) respectively, p means each pixel location and superscript $*$ denotes the groundtruth. The default settings are: w_{2D}=0.1, $w_{\Delta Z}$=5, w_{RZ}=10.

1.2 Running time and memory

Table 5 provides detailed information about running time and memory of the state-of-the-art top-down method [22] and our method. The input size of images is fixed, 832×512. Note that our method is almost not affected by the number of people in the image.

Table 5: Running time and memory comparison.

		3-people		20-people	
		Time(ms)	Memory(M)	Time(ms)	Memory(M)
[22]	DetectNet	120.0	899	120.0	899
	PoseNet	14.7	815	71.8	1491
	RootNet	13.0	803	58.9	1051
Ours	SMAP	57.0	1379	57.0	1379
	DAPA	4.5	-	8.8	-
	RefineNet	0.80	∼0.5	0.83	∼0.5

2 More results compared with SOTA

Due to the limited space, only the average PCK_{abs} is reported in the main manuscript. Here we provide more thorough experimental results. Table 6 presents sequence-wise PCK_{abs} on the MuPoTS-3D dataset and demonstrates that our PCK_{abs} is higher than the state-of-the-art top-down method [22], especially for outdoor scenarios (TS6-TS20). Table 7 shows that our model has higher PCK_{rel} compared with all bottom-up methods and most top-down methods except [22]. Note that we have higher AUC_{rel} compared with [22] as we state in the main manuscript. Table 8 shows the results on the Human3.6M dataset.

3 More ablation analysis

3.1 Effect of the multi-task structure

SMAP simultaneously output 2D information (keypoint heatmaps and PAFs), root depth map, and part relative-depth map. To analyze the impact of our single-shot multi-task structure on root localization, we delete some of the output branches and evaluate the performance, as indicated in Table 9. One variant is only to regress the root position and its depth alone (row 2 of Table 9). This variant can obtain an acceptable result, which reflects the significance of our bottom-up design for root localization. Another variant which adds the keypoint heatmaps and PAFs branches (row 3 of Table 9) significantly improves the performance, indicating that 2D cues (pose, body size) are also beneficial to root depth estimation. Nevertheless, this variant is still inferior to the full model.

3.2 Influence of camera intrinsics

Here we make three comparisons:1) full model with known camera intrinsics. 2) full model without camera intrinsics. 3) without normalization.

RtError of our full model reaches 23.3cm on the MuPoTS-3D dataset. If the intrinsic parameter is not provided (use default intrinsics), RtError increases to 67cm. Note that the ordinal depth relation remains unchanged. If the model lacks normalization, RtError is as high as 120cm.

Table 6: Sequence-wise PCK_{abs} on the MuPoTS-3D dataset for matched groundtruths.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
Moon et al. [22]	**59.5**	**45.3**	**51.4**	**46.2**	53.0	**27.4**	23.7	**26.4**	**39.1**	23.6	
Ours	42.1	41.4	46.5	16.3	**53.0**	26.4	**47.5**	18.7	36.7	**73.5**	
	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg.
Moon et al. [22]	18.3	14.9	**38.2**	29.5	36.8	23.6	14.4	20.0	18.8	25.4	31.8
Ours	**46.0**	**22.7**	24.3	**38.9**	**47.5**	**34.2**	**35.0**	**20.0**	**38.7**	**64.8**	**38.7**

Table 7: PCK_{rel} on the MuPoTS-3D dataset for matched groundtruths. 'T' denotes top-down methods while 'B' denotes bottom-up methods.

		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
Rogez et al. [33]	T	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1	
Rogez et al. [34]	T	88.0	73.3	67.9	74.6	81.8	50.1	60.6	60.8	78.2	89.5	
Dabral et al. [6]	T	85.8	73.6	61.1	55.7	77.9	53.3	75.1	65.5	54.2	81.3	
Moon et al. [22]	T	**94.4**	78.6	**79.0**	**82.1**	86.6	**72.8**	81.9	75.8	**90.2**	**90.4**	
Mehta et al. [20]	B	81.0	64.3	64.6	63.7	73.8	30.3	65.1	60.7	64.1	83.9	
Mehta et al. [19]	B	88.4	70.4	68.3	73.6	82.4	46.4	66.1	**83.4**	75.1	82.4	
Ours	B	89.9	**88.3**	78.9	78.2	**87.6**	51.0	**88.5**	71.6	70.3	89.2	
		S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg.
Rogez et al. [33]	T	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
Rogez et al. [34]	T	70.8	74.4	72.8	64.5	74.2	84.9	85.2	78.4	75.8	74.4	74.0
Dabral et al. [6]	T	**82.2**	71.0	70.1	67.7	69.9	90.5	85.7	**86.3**	85.0	**91.4**	74.2
Moon et al. [22]	T	79.4	79.9	**75.3**	**81.0**	**81.0**	90.7	89.6	83.1	81.7	77.3	**82.5**
Mehta et al. [20]	B	71.5	69.6	69.0	69.6	71.1	82.9	79.6	72.2	76.2	85.9	69.8
Mehta et al. [19]	B	76.5	73.0	72.4	73.8	74.0	83.6	84.3	73.9	**85.7**	90.6	75.8
Ours	B	76.3	**82.0**	70.8	65.2	80.4	**91.6**	**90.4**	83.4	84.3	91.2	80.5

Table 8: MPJPE Results on Human3.6M dataset. Note that there is no groundtruth bounding box information in inference time.

Method	MPJPE
Rogez et al. [33]	87.7
Mehta et al. [20]	69.9
Dabral et al. [6]	65.2
Mehta et al. [19]	63.6
Rogez et al. [34]	63.5
Moon et al. [22]	54.4
Ours	**54.1**

Table 9: Ablation study of the structure design on the MuPoTS-3D dataset. Note that our full model consists of root depth, relative depth and 2D branches.

Design	Recall	PCK_{root}	PCK_{abs}	PCK_{rel}	PCOD
Full Model	**92.3**	**45.5**	**38.7**	**80.5**	**97.0**
Root Depth Only	85.0	29.9	-	-	88.3
Root Depth + 2D Branches	92.1	43.6	-	-	96.7