Supplementary Material of "Camera Distance-aware Top-down Approach for 3D Multi-person Pose Estimation from a Single RGB Image"

Gyeongsik Moon
Department of ECE, ASRI
Seoul National University
mks0601@snu.ac.kr

Ju Yong Chang
Department of EI
Kwangwoon University
juyong.chang@gmail.com

Kyoung Mu Lee
Department of ECE, ASRI
Seoul National University
kyoungmu@snu.ac.kr

In this supplementary material, we present more experimental results that could not be included in the main manuscript due to the lack of space.

1. Derivation of Equation 1

We provide a derivation of Equation 1 of the main manuscript with reference to Figure A ,which shows a pinhole camera model. The green and blue arrows represent the human root joint centered x and y-axes, respectively. The yellow lines show rays, and c is the hole. d, f, and l_{sensor} are distance between camera and the human root joint (mm), focal length (mm), and the length of human on the image sensor (mm), respectively.

According to the definition of $\tan$,

$$\tan\theta_x = \frac{0.5l_{x,real}}{d} = \frac{0.5l_{x,sensor}}{f},$$

Let p_x be per pixel distance factor in x-axis. Then,

$$d = f\frac{l_{x,real}}{l_{x,sensor}} = fp_x\frac{l_{x,real}}{l_{x,sensor}p_x} = \alpha_x\frac{l_{x,real}}{l_{x,img}},$$

Above equations are also valid in y-axis. Therefore,

$$d = f\frac{l_{y,real}}{l_{y,sensor}} = fp_y\frac{l_{y,real}}{l_{y,sensor}p_y} = \alpha_y\frac{l_{y,real}}{l_{y,img}},$$

Finally,

$$d = \sqrt{\alpha_x\alpha_y\frac{l_{x,real}}{l_{x,img}}\frac{l_{y,real}}{l_{y,img}}} = \sqrt{\alpha_x\alpha_y\frac{A_{real}}{A_{img}}}.$$

2. Comparison of 3D human root localization with previous approaches

We compare previous absolute 3D human root localization methods [5, 8] with the proposed RootNet on the Human3.6M dataset [2] based on protocol 2.

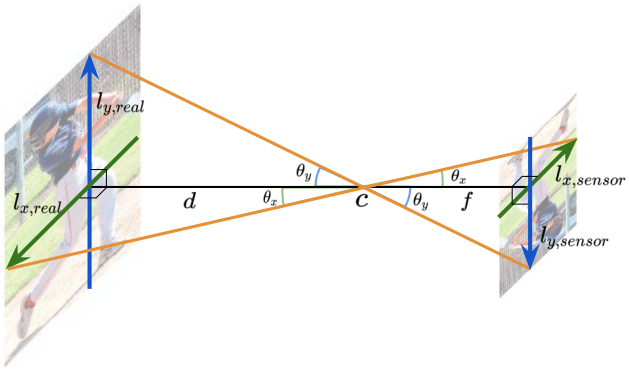

Figure A: Visualization of a pinhole camera model.

Methods	MRPE	MRPE$_x$	MRPE$_y$	MRPE$_z$
Baseline [5, 8]	267.8	27.5	28.3	261.9
W/o limb joints	226.2	24.5	24.9	220.2
RANSAC	213.1	24.3	24.3	207.1
RootNet (Ours)	**120.0**	**23.3**	**23.0**	**108.1**

Table A: MRPE comparisons between previous distance minimization-based approaches [5, 8] and our RootNet on the Human3.6M dataset. MRPE$_x$, MRPE$_y$, and MRPE$_z$ represent the mean of the errors in the x, y, and z axes, respectively.

DetectNet	RootNet	PoseNet	**Total**
0.120	0.010	0.011	**0.141**

Table B: Seconds per frame for each component of our framework.

Previous approaches [5, 8] simultaneously estimate 2D image coordinates and 3D camera-centered root-relative coordinates of keypoints. Then, absolute camera-centered coordinates of the human root are obtained by minimizing the distance between 2D predictions and projected 3D predictions. For optimization, linear least-squares formulation is used. To measure the errors of their method, we imple-

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 groundtruths																					
Ours	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
Accuracy only for matched groundtruths																					
Ours	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

Table C: Sequence-wise 3DPCK$_{abs}$ on the MuPoTS-3D dataset.

Methods	Hd.	Nck.	Sho.	Elb.	Wri.	Hip	Kn.	Ank.	Avg
Ours	37.3	35.3	33.7	33.8	30.4	30.3	31.0	25.0	31.5

Table D: Joint-wise 3DPCK$_{abs}$ on the MuPoTS-3D dataset. All groundtruths are used for evaluation.

mented and used ResNet-152-based model of Sun *et al*. [9] as a 2D pose estimator and model of Martinez *et al*. [4] as a 3D pose estimator, which are state-of-the-art methods. In addition, to minimize the effect of outliers in 3D-to-2D fitting, we excluded limb joints when fitting. Also, we performed RANSAC with a various number of joints to get optimal joint set for fitting instead of using heuristically selected joint set.

Table A shows our RootNet significantly outperforms previous approaches. Furthermore, the RootNet can be designed independently of the PoseNet, giving design flexibility to both models. In contrast, the previous 3D root localization methods [5,8] require both of 2D and 3D predictions for the root localization, which results in lack of generalizability.

3. Running time of the proposed framework

In Table B, we report seconds per frame for each component of our framework. The running time is measured using a single TitanX Maxwell GPU. As the table shows, most of the running time is consumed by DetectNet. It is hard to directly compare running time with previous works [5, 8] because they did not report it. However, we guess that there would be no big difference because models of [8] and [5] are similar with [7] and [1] whose speed is 0.2 and 0.11 seconds per frame, respectively.

4. Absolute 3D multi-person pose estimation errors

For the continual study of the 3D multi-person pose estimation, we report 3DPCK$_{abs}$ in Table C and D. As previous works [3, 6] did not report 3DPCK$_{abs}$, we only report our result.

5. Qualitative results

Figures B and C show qualitative results of our 3D multi-person pose estimation framework on the MuPoTS-3D [6] and COCO [3] datasets, respectively. Note that COCO dataset consists of *in-the-wild* images which are hardly included in the 3D human pose estimation training sets [2, 6].

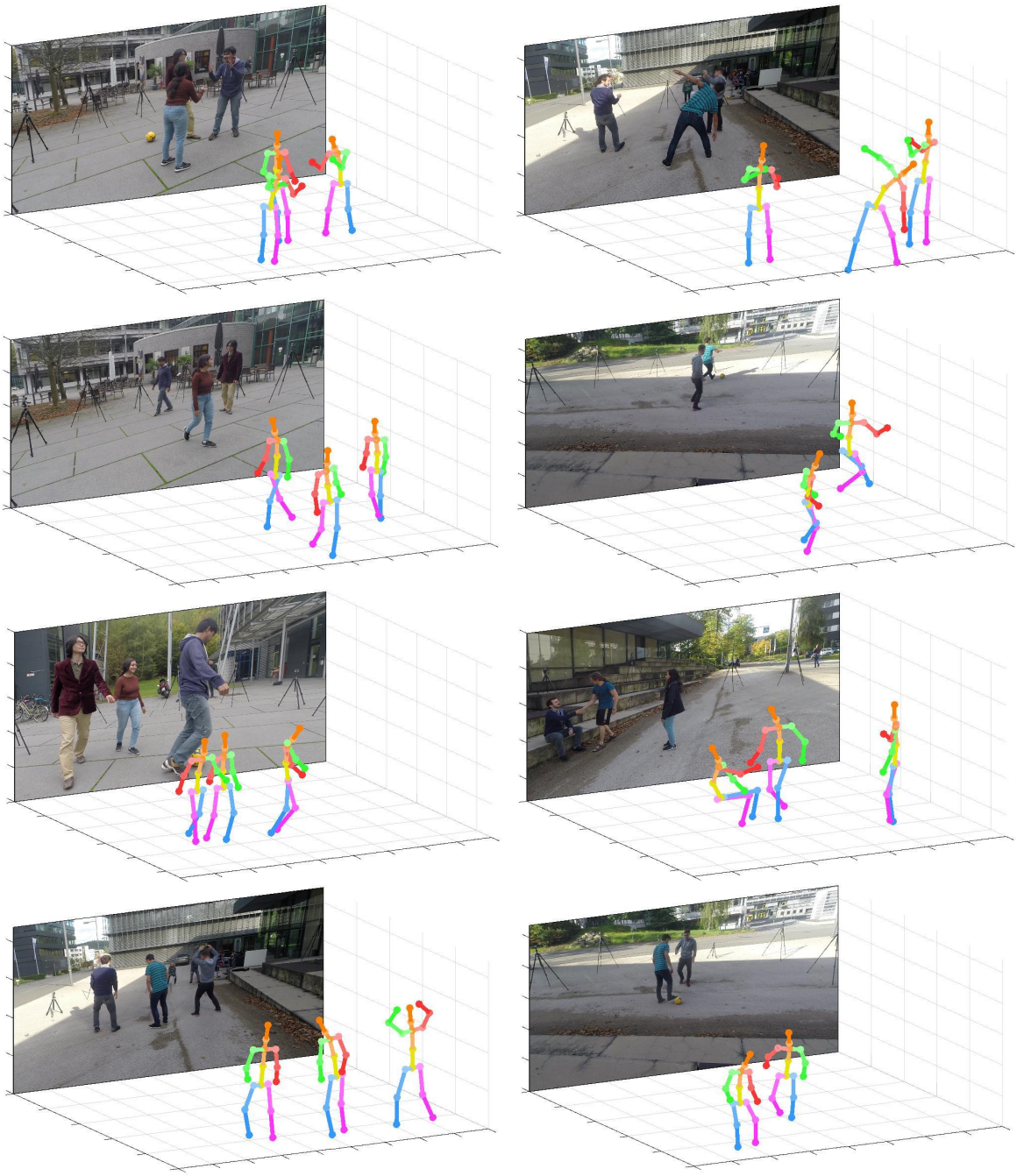

Figure B: Qualitative results of applying our method on the MuPoTS-3D dataset [6].

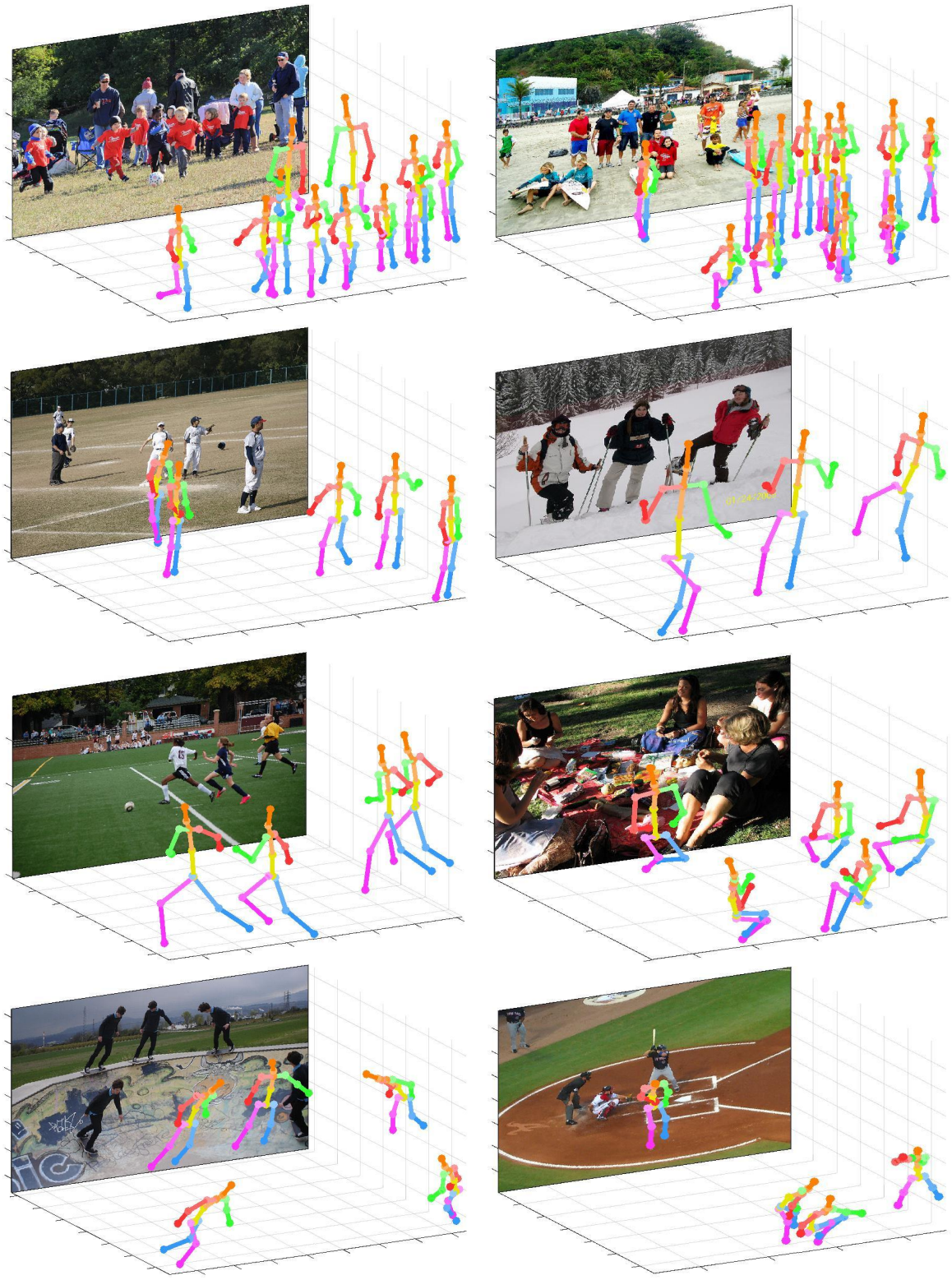

Figure C: Qualitative results of applying our method on the COCO 2017 [3] validation set.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*, 2017.

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2014.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[4] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[5] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017.

[6] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[8] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017.

[9] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018.