Supplementary Material for
Single-Stage is Enough: Multi-Person Absolute 3D Pose Estimation

Abstract

In this supplementary material, we present fully detailed information on 1) Quantitative comparisons of MPJPE on CMU Panoptic [5] dataset; 2) hyper-parameters analysis for loss function; 3) sequence-wise results on MuPOTS-3D [8] dataset; 4) qualitative results on MuPOTS-3D [8] dataset; 5) multi-view visualization 3D poses upon in-the-wild images from COCO [6] validation set.

A This appendix provides quantitative comparisons between state-of-the-art methods and ours on CMU Panoptic [5] dataset.

MPJPE measures the accuracy of the 3D root-relative pose. It is computed by using the Euclidean distance between the estimated 3D joints and the groundtruth positions. Quantitative comparisons between state-of-the-art methods and ours are provided in Tab. 1.

Table 1. Quantitative comparisons of MPJPE on CMU Panoptic [5]. T denotes Top-down method, and B denotes Bottom-up method.

Methods	Haggling	Mafia	Ultim.	Piazza	Mean↓
CRP [4] (T)	129.6	133.5	153.0	156.7	143.2
CDMP [10] (T)	89.6	91.3	79.6	90.1	87.6
HMOR [13] (T)	50.9	50.5	50.7	68.2	51.6
PandaNet [1] (T)	**40.6**	**37.6**	**31.3**	**55.8**	**42.7**
MubyNet [14] (B)	72.4	78.8	66.8	94.3	72.1
SMAP [15] (B)	63.1	**60.3**	**56.6**	**67.1**	**61.8**
LoCO [3] (B)	**45.0**	95.0	58.0	79.0	69.0
DRM (Ours)	**52.5**	**56.1**	**47.5**	**70.3**	**56.7**

B This appendix provides experiments on parameter analysis of our loss function.

In sec. 3.3 in the manuscript, we use the hyper-parameters λ_o, λ_{rz}, $\lambda_{\Delta z}$ and λ_p to balance different loss items. We compare the performance of our model with different hyper-parameters. On the whole, we should choose better values for loss hyper-parameters to solve the unbalance problem between all loss items.

Table 2. Study of loss hyper-parameters on the MuPoTS-3D [8] dataset for matched groundtruths.

Loss hyper-parameters	PCK_{rel}	PCK_{abs}	PCK_{root}	AUL_{rel}
λ_o, λ_{rz}, $\lambda_{\Delta z}$=0.3 λ_p=0.03	83.5	39.8	43.4	42.9
λ_o, λ_{rz}, $\lambda_{\Delta z}$=0.03 λ_p=0.03	83.6	39.7	43.7	43.0
λ_o, λ_{rz}, $\lambda_{\Delta z}$=0.03 λ_p=0.003	**85.1**	**41.0**	**45.6**	**45.4**
λ_o, λ_{rz}, $\lambda_{\Delta z}$=0.03 λ_p=0.0003	84.9	39.3	42.3	45.2

C This appendix provides more thorough experiments, *i.e.*, sequence-wise results on the MuPoTS-3D [8] dataset.

Due to the limited space, only the average PCK_{abs} and PCK_{rel} are reported in the main manuscript. Here we provide more detailed experimental results. Tab. 3 provides sequence-wise PCK_{abs} on the MuPoTS-3D [8] dataset and demonstrates that over half sequences of our PCK_{abs} is higher than the state-of-the-art bottom-up method SMAP [15]. Tab. 4 shows that our model has higher PCK_{rel} in most sequences compared with all bottom-up methods and top-down methods.

D This appendix provides an additional visualized results of outdoor images from MuPoTS-3D [8] testing set. (Sec. 4.2)

Fig. 1 gives the visualized results of the estimated 3D poses upon outdoor images from MuPoTS-3D [8] testing set. It is shown that in outdoor challenging scenario (containing scale variance, occlusion, and drastic illumination changes), our method still performs surprisingly well.

E This appendix provides an additional results of the estimated 3D poses upon in-the-wild images from COCO [6] validation set in three views.

Fig. 2 provides the visualization results of the estimated 3D poses from in-the-wild images in three views, containing several scenarios, *e.g.*, various poses, scale variance, occlusion, *etc.* It can be seen from the top-down view that our method performs well in estimating the depth of all instances, which is hard to capture from other views.

Table 3. Sequence-wise PCK_{abs} on the MuPoTS-3D [8] dataset for matched groundtruths.

Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
CDMP [10] (Top-down)	**59.5**	45.3	51.4	46.2	53.0	27.4	23.7	**26.4**	**39.1**	23.6	
SMAP [15] (Bottom-up)	42.1	41.4	46.5	16.3	53.0	26.4	47.5	18.7	36.7	**73.5**	
DRM (Ours, single-stage)	57.8	**51.8**	**51.8**	**54.3**	**61.1**	**49.5**	**41.5**	9.4	33.4	73.0	
	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg.
CDMP [10] (Top-down)	18.3	14.9	**38.2**	29.5	36.8	23.6	14.4	20.0	18.8	25.4	31.8
SMAP [15] (Bottom-up)	**46.0**	22.7	24.3	38.9	**47.5**	**34.2**	**35.0**	20.0	38.7	**64.8**	38.7
DRM (Ours, single-stage)	18.4	**50.0**	25.1	**40.5**	43.9	25.8	34.1	**21.4**	**40.5**	36.2	**41.0**

Table 4. Sequence-wise PCK_{rel} on the MuPoTS-3D [8] dataset for matched groundtruths.

Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
LCR-Net [11] (Top-down)	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1	
LCR-Net++ [12] (Top-down)	87.3	61.9	67.9	74.6	78.8	48.9	58.3	59.7	78.1	89.5	
HG-RCNN [2] (Top-down)	85.1	67.9	73.5	76.2	74.9	52.5	65.7	63.6	56.3	77.8	
CDMP [10] (Top-down)	**94.4**	77.5	79.0	81.9	85.3	72.8	81.9	75.7	**90.2**	**90.4**	
ORPM [7] (Bottom-up)	81.0	59.9	64.4	62.8	68.0	30.3	65.0	59.2	64.1	83.9	
Xnect [9] (Bottom-up)	88.4	65.1	68.2	72.5	76.2	46.2	65.8	64.1	75.1	82.4	
SMAP [15] (Bottom-up)	88.8	71.2	77.4	77.7	80.6	49.9	**86.6**	51.3	70.3	89.2	
DRM (Ours,single-stage)	91.2	**81.0**	**83.8**	**84.2**	**90.6**	**77.3**	80.1	**81.7**	88.9	86.5	
	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg.
LCR-Net [11] (Top-down)	50.2	51.0	51.6	49.3	56.2	66.5	65.2	62.9	66.1	59.1	53.8
LCR-Net++ [12] (Top-down)	69.2	73.8	66.2	56.0	74.1	82.1	78.1	72.6	73.1	61.0	70.6
HG-RCNN [2] (Top-down)	76.4	70.1	65.3	51.7	69.5	87.0	82.1	80.3	78.5	70.7	71.3
CDMP [10] (Top-down)	79.2	79.9	75.1	72.7	81.1	89.9	**89.6**	81.8	81.7	76.2	81.8
ORPM [7] (Bottom-up)	67.2	68.3	60.6	56.5	69.9	79.4	79.6	66.1	66.3	63.5	65.0
Xnect [9] (Bottom-up)	74.1	72.4	64.4	58.8	73.7	80.4	84.3	67.2	74.3	67.8	70.4
SMAP [15] (Bottom-up)	72.3	81.7	63.6	44.8	79.7	86.9	81.0	75.2	73.6	67.2	73.5
DRM (Ours,single-stage)	**82.9**	**87.3**	**82.7**	**76.1**	**84.4**	**92.3**	88.1	**85.6**	**85.6**	**92.4**	**85.1**

References

[1] A. Benzine, F. Chabot, B. Luvison, Q. C. Pham, and C. Achard. Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In *CVPR*, 2020. 1

[2] Rishabh Dabral, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3d human pose estimation from monocular images. In *3DV*, 2019. 2

[3] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *CVPR*, 2020. 1

[4] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 1

[5] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, and I. Matthews. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, pages 1–1, 2016. 1

[6] T. Y. Lin, M. Maire, S. Belongie, J. Hays, and C. L. Zitnick. Microsoft coco: Common objects in context. *ECCV*, 2014. 1

[7] 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. 2

[8] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, and C. Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018. 1, 2, 3

[9] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, and Christian Theobalt. Xnect: real-time multi-person 3d motion capture with a single rgb camera. *TOG*, 39(4), 2020. 2

[10] G. Moon, J. Y. Chang, and K. M. Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, 2020. 1, 2

[11] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 2

[12] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *TPAMI*, 2019. 2

[13] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, pages 242–259, 2020. 1

[14] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *NIPS*, pages 8410–8419, 2018. 1

[15] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *ECCV*, pages 550–566, 2020. 1, 2

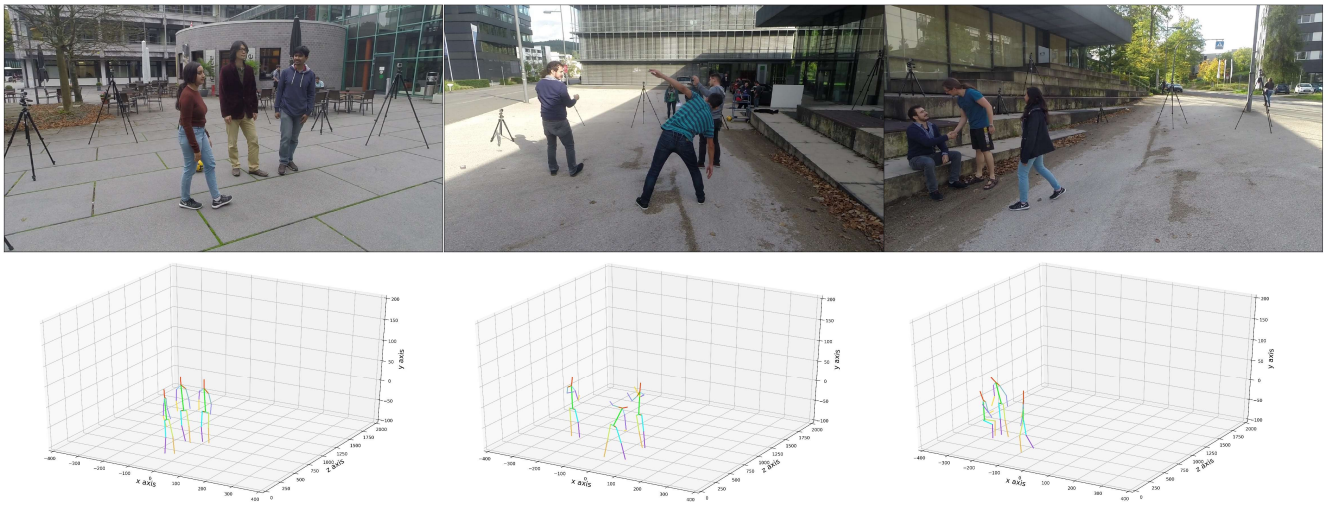
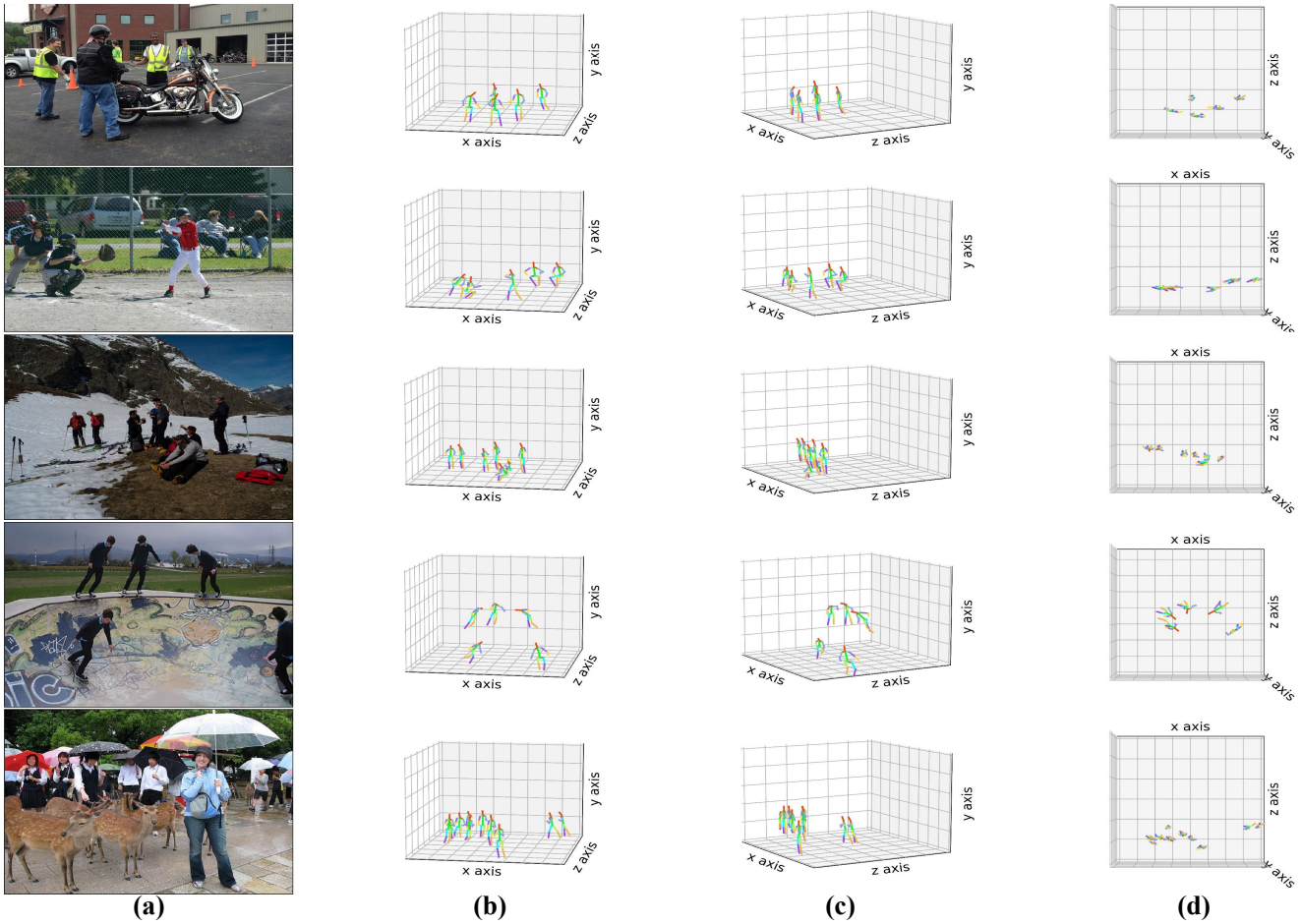


Figure 1. Visualized results of outdoor images from MuPoTS-3D [8] testing set. Top row: input image. Bottom row: corresponding multi-person 3D pose estimation of the proposed DRM.



(a) **(b)** **(c)** **(d)**

Figure 2. Visualization results of different views for in-the-wild images from COCO validation set. (a) origin image (b) front view (c) right view (d) top-down view