Body Meshes as Points (Supplementary File)

Jianfeng Zhang[1] Dongdong Yu[2] Jun Hao Liew[1] Xuecheng Nie[3] Jiashi Feng[1]
[1]National University of Singapore [2]ByteDance AI Lab [3]Yitu Technology
{zhangjianfeng,liewjunhao}@u.nus.edu yudongdong@bytedance.com elefjia@nus.edu.sg

This supplementary file includes additional details that were not included in the main manuscript due to space limits. We start with more implementation details (Appendix A). Then, we provide ablation study on the pyramid level K used in our BMP model (Appendix B). Finally, we provide more qualitative results, including successes, failures and comparisons with the baseline models (Appendix C).

A. Implementation details

In our preliminary experiments, we observe that the direct end-to-end training of the whole network from scratch cannot achieve the best performance. We argue that this is likely because the tasks of person localization and body mesh recovery perform differently during the training process, *i.e*. the body mesh branch needs more training iterations than the localization branch. Therefore, we utilize a multi-stage training scheme, which is more stable and effective in practice. More specifically, we first pretrained the body mesh branch with cropped images from single-person samples from Human3.6M [2], MPI-INF-3DHP [7], COCO [5], LSP [3], LSP Extended [4] and MPII [1] for roughly 80 epochs. Then we trained the whole network end-to-end on full-images with multiple persons for 30 epochs. Our proposed ordinal depth loss is active in the second stage. We trained our BMP model in 4 V100 GPUs with a learning rate of $1e-4$ using the Rectified Adam optimizer [6] in the pretraining stage and $1e-5$ in the full training stage.

B. Ablation study on pyramid level K

The default BMP model uses five pyramids to localize and estimate body meshes of person instances with different scales (Table 1). From P2 to P6, the corresponding grid numbers are $[40, 36, 24, 16, 12]$, respectively. We aim to evaluate the impact of the number of pyramid level K used in the model. Specifically, we compare the performance of models using different pyramid level K on three multi-person datasets. Results are shown in Table S1. $K = 1$ denotes the model trained using only a single-scale pyramid level with $G = 24$; $K = 3$ represents the model trained using three pyramid levels with the grid numbers $[36, 24, 12]$; and $K = 5$ is our BMP model trained using five pyramid levels with the grid numbers $[40, 36, 24, 16, 12]$. From the results, we can observe using $K = 1$ can already achieve 147.3 mm MPJPE on the challenging Panoptic dataset, outperforming previous two-stage methods [8, 9]. This clearly verifies the effectiveness of our model's two parallel branch design. When using more pyramid levels in our model, the performance can be largely improved in all datasets, which demonstrates the benefits of using an additional depth dimension to represent person instances.

K	Panoptic (↓)	3DPW (↓)	MuPoTS-3D (↑)
1	148.7	112.4	68.32
3	142.3	106.2	71.52
5	135.4	104.1	73.83

Table S1. **Ablation for pyramid leve K.** We report MPJPE for Panoptic and 3DPW, and 3DPCK for MuPoTS-3D.

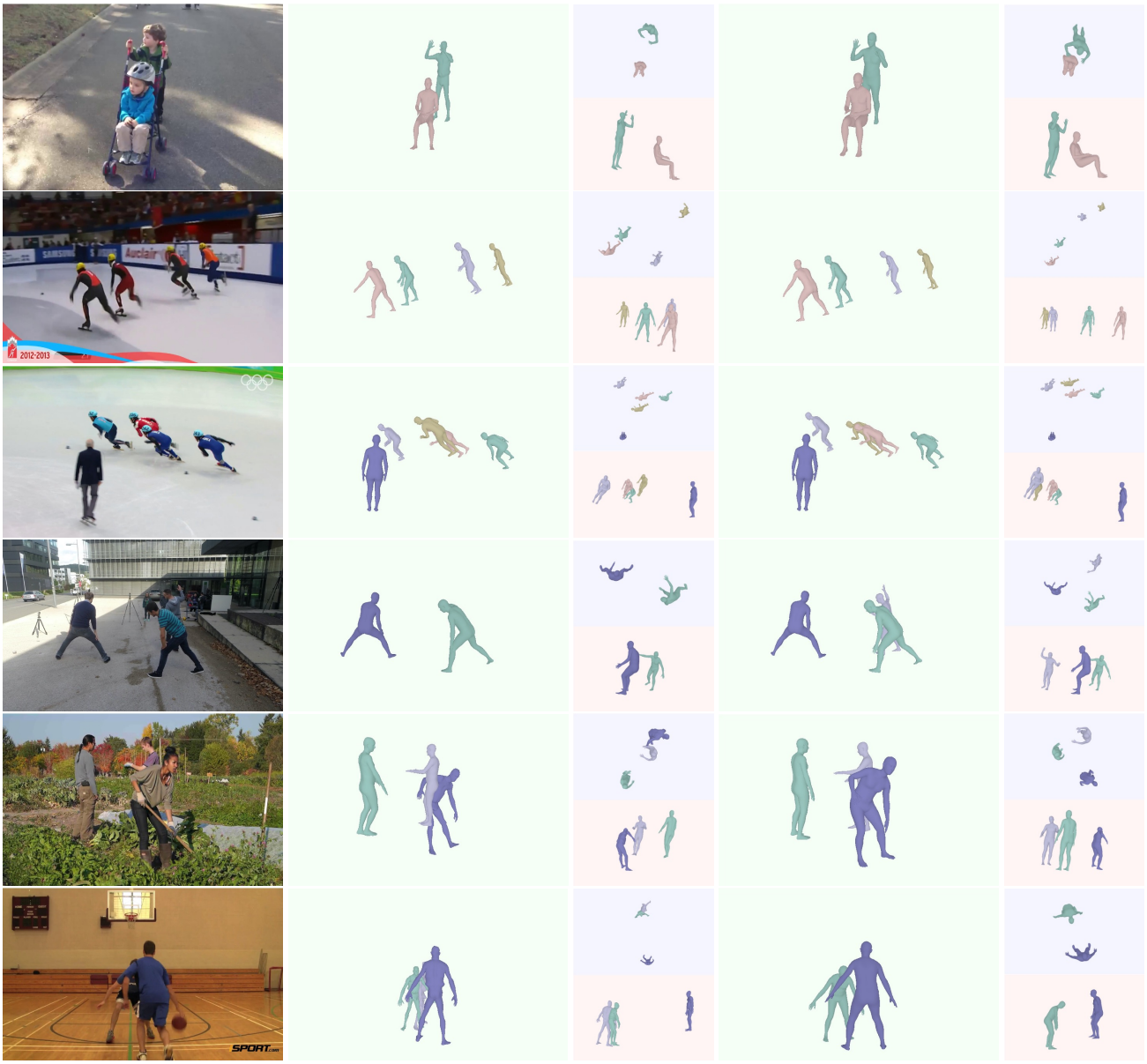
C. More qualitative results

For our qualitative evaluation, we first provide more comparisons between our baseline model and our BMP model trained with our proposed methods in Fig. S1. Then we provide more successful reconstructions from the datasets we use in our evaluation in Fig. S2 and Fig. S3, respectively. Finally, in Fig. S4, we present some representative failure cases of our BMP model.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[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. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.



Input image Baseline Ours

Figure S1. **Qualitative effect of proposed method.** Results of baseline 1 (BMP trained w/o L_{rank}) (middle 1st to 3rd rows), baseline 2 (BMP using 2D spatial representation) (middle 4th row), baseline 3 (BMP trained w/o occlusion augmentation) (middle 5th and 6th rows) and BMP (right). As expected, the proposed methods take effect on producing better results (*i.e.*, more depth-coherent reconstructions and robust to overlapping instances and partial observations).

[3] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[4] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[5] 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.

[6] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv*, 2019.

[7] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. on Graphics*, 36(4):44, 2017.

[8] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018.

Figure S2. **Successful results.** We visualize the reconstructions of our BMP model from different viewpoints.

[9] 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 *NeurIPS*, 2018.

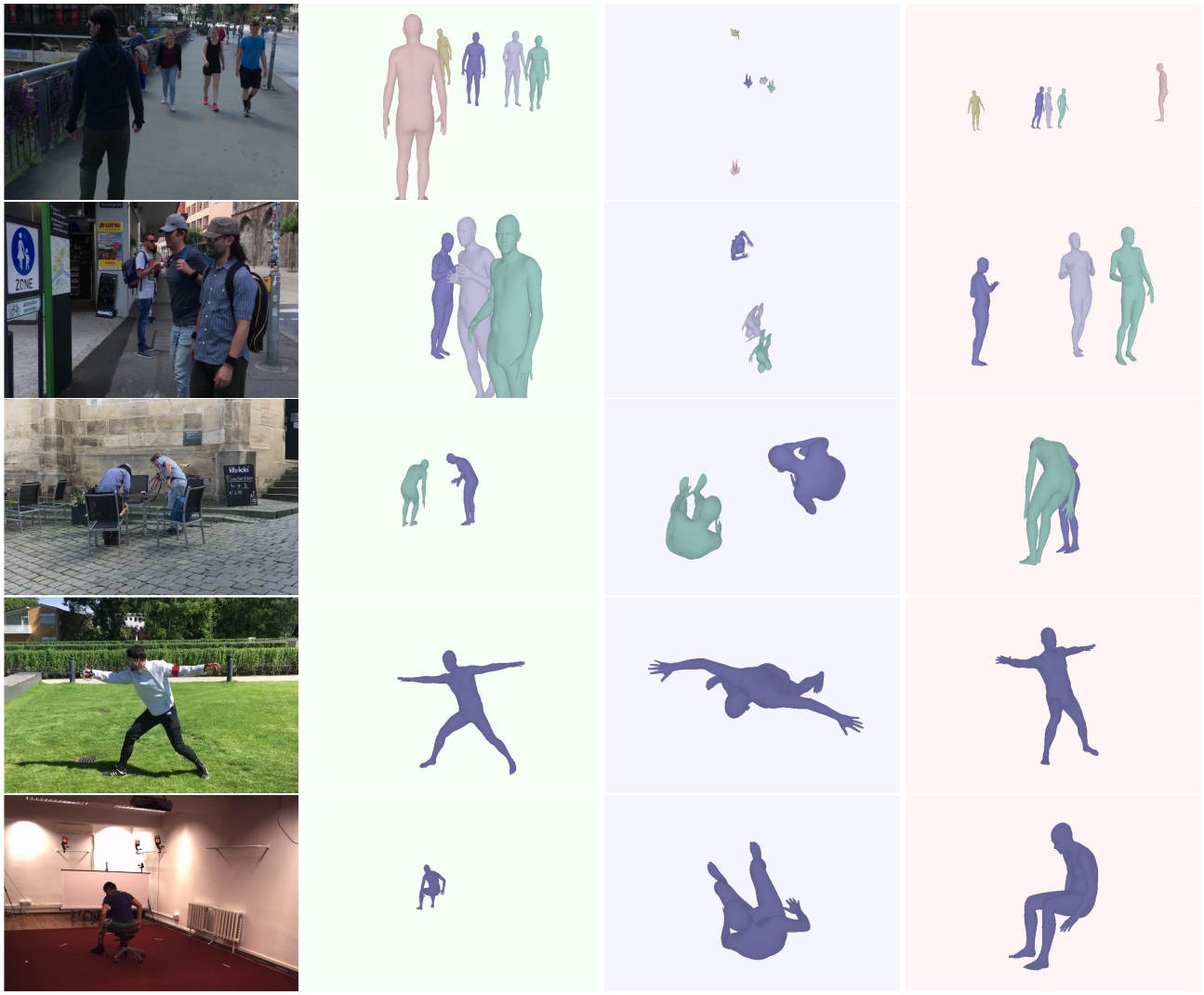


Figure S3. **Successful results.** We visualize the reconstructions of our BMP model from different viewpoints.

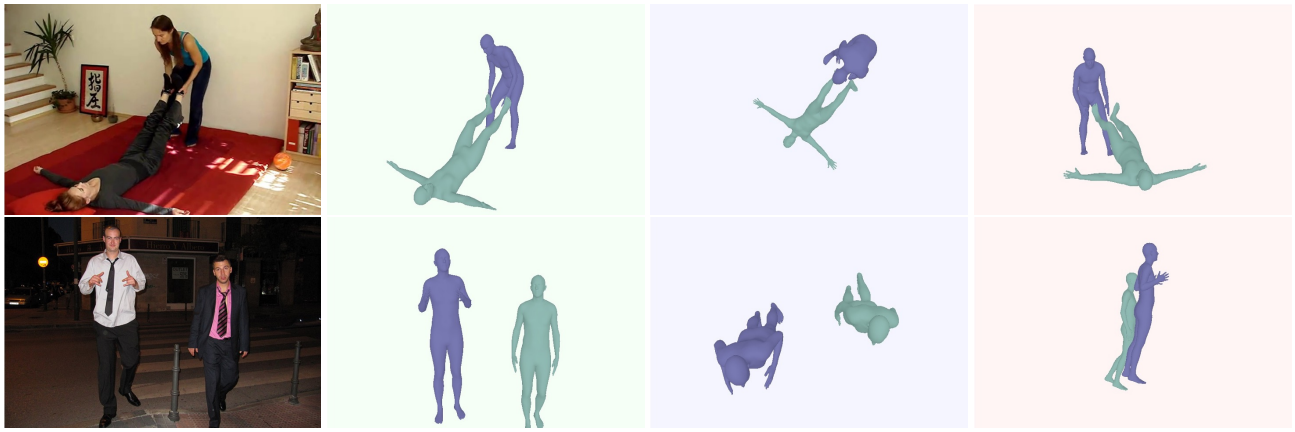


Figure S4. **Failure cases.** We visualize the reconstructions of our BMP model from different viewpoints. For the first image, BMP estimates the relative depth ordering correctly, but overestimates the distance between the two people, who are in contact. This motivates us to explore how to extend BMP to inter-person interactions modeling in the future. For the second image, BMP estimates the position of the person on the right to be farther away from the camera than the person in the left; while actually the two people stand with roughly the same depth in the image.