Supplementary Material *for*
"Learning to Estimate Robust 3D Human Mesh from In-the-Wild Crowded Scenes"

Hongsuk Choi Gyeongsik Moon JoonKyu Park Kyoung Mu Lee

ECE & ASRI, Seoul National University, Korea
{redarknight,mks0601,jkpark0825,kyoungmu}@snu.ac.kr

method	training datasets	MPJPE↓	PA-MPJPE↓	MPVPE↓
SPIN [11]	Human3.6M [6], MPI-INF-3DHP [15], MSCOCO [13], MPII [2], LSP [8], LSP-Extended [9]	121.2	69.9	144.1
Pose2Mesh [5]	Human3.6M [6], MuCo-3DHP [16], MSCOCO [13]	124.8	79.8	149.5
I2L-MeshNet [17]	Human3.6M [6], MuCo-3DHP [16], MSCOCO [13]	115.7	73.5	162.0
ROMP [19]	Human3.6M [6], MPI-INF-3DHP [15], MSCOCO [13], MPII [2], LSP [8], LSP-Extended [9], AICH [21], MuCo-3DHP [16], OH [23], PoseTrack [1], CrowdPose [12]	104.8	63.9	127.8
3DCrowdNet (Ours)	**MuCo-3DHP [16], MSCOCO [13]**	**88.3**	**59.2**	**112.8**

Table A. Comparison on 3DPW-Crowd between 3DCrowdNet and previous methods. 3DCrowdNet uses the least training datasets and achieves the best accuracy on in-the-wild crowded scenes.

In this supplementary material, we first clarify that the best accuracy of 3DCrowdNet in the main manuscript's Table 5 is not from using more training data in Section A.1. Then, we go through the details of testing sets in Section A.2. In Section B, we provide additional qualitative results, including comparisons with other methods. In Section C, we discuss the limitation of 3DCrowdNet and its potential solutions. Last, we give details of 2D pose estimators used in our experiments in Section D.

A. Datasets

A.1. Training sets of different methods

Table A demonstrates that the superiority of 3DCrowdNet does not come from using more training data. It shows the training datasets used in the previous methods of the main manuscript's Table 5. We trained 3DCrowdNet on one MoCap dataset and one in-the-wild 2D dataset, which is the least training set among methods, and we tested it on 3DPW-Crowd. It still significantly outperforms the previous methods in all metrics. We used 2D pose outputs of HigherHRNet [4], which is trained only on MSCOCO [13]. The results strongly support that our contributions listed in the main manuscript's Section 1.

A.2. Details of testing sets

3DPW-Crowd. The sequence names of 3DPW-Crowd are *courtyard_hug_00* and *courtyard_dancing_00*, a subset of the 3DPW [20] validation set. 3DPW-Crowd contains 1073 images and 1923 persons with GT 3D pose and shape annotations. The average bounding box IoU is 37.5%, and the CrowdIndex [12] is 49.3%. We used 14 joints defined by Human3.6M [6] for evaluating PA-MPJPE and MPJPE following the previous works [5, 11, 17].

MuPoTS. MuPoTS [16] contains 20 sequences, 8370 images, and 20899 persons with GT 3D pose annotations. The sequences are captured indoors and outdoors, and GT 3D poses are obtained by a multi-view marker-less motion capture system. The average bounding box IoU is 3.8%, and the CrowdIndex [12] is 13.2%. We used the official MATLAB code for evaluation.

CMU-Panoptic. We selected four sequences that show people doing social activities, namely *Haggling*, *Mafia*, *Ultimatum*, and *Pizza* following [7, 22]. Sequences captured

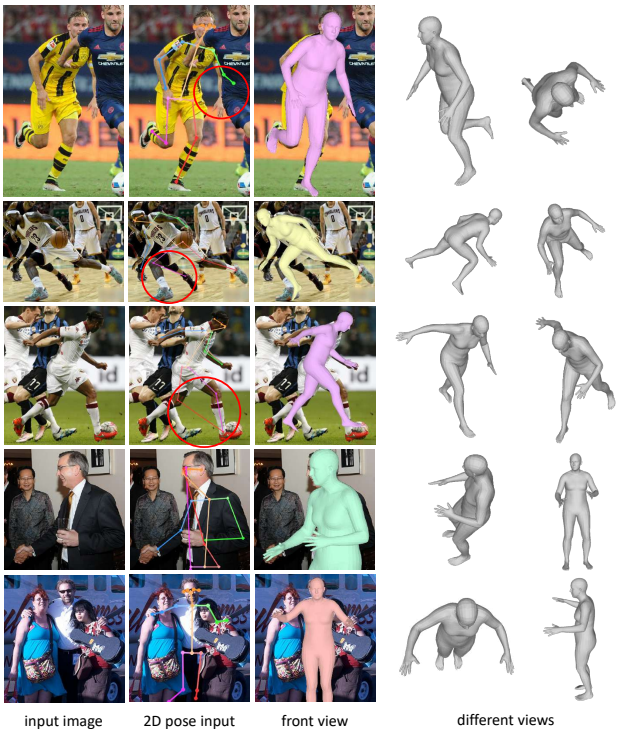

Figure A. Visualization of 3D meshes from different viewpoints. Our 3DCrowdNet can recover a 3D shape that best describes the target person in an image, even when provided with inaccurate 2D poses, using the target person's image features.

by the 16th and 30th cameras are selected. The sequences contain 9600 frames and 21,404 persons with GT 3D pose annotations. The average bounding box IoU is 2.0%, and the CrowdIndex [12] is 11.1%. We used pre-processed GT annotations and followed the evaluation protocol of [7] in their official code repository.

3DPW. We used the test set of 3DPW [20] following the official split protocol. The test set contains 26240 images and 35515 persons with GT 3D pose and shape annotations. The average bounding box IoU is 3.7%, and the CrowdIndex [12] is 4.9%. Sequences starring one actor are excluded in computing the bounding box IoU and the CrowdIndex. We used 14 joints defined by Human3.6M [6] for evaluating PA-MPJPE and MPJPE following the previous works [5, 11, 17].

B. More qualitative results

Accurate 3D meshes from erroneous 2D pose input. Figure A shows that our 3DCrowdNet can estimate robust 3D meshes, given inaccurate 2D poses from in-the-wild crowded scenes. Due to inter-person occlusion and overlapping bounding boxes between people, 2D pose estimators [3, 4] may produce inaccurate joint predictions as shown in the first, second, and third rows. To handle such cases, the feature extractor of 3DCrowdNet assigns don't-care values to the joint predictions with low confidence (*e.g.* lower than 0.1 for outputs from [4]) using heatmap representation, as discussed in Section 3.1 of the main manuscript. Then, the joint-based regressor of 3DCrowdNet refines the 2D pose heatmap, while predicting a 3D pose with image features containing the image's context information. The 2D pose heatmap has a different joint set, a superset of joint sets defined by multiple datasets, with the 3D pose's joint set, an intersection of joint sets defined by multiple datasets. Figure B depicts each joint set. Last, the joint-based regressor samples image features using the (x,y) pixel positions of the 3D pose and estimates human model parameters, SMPL [14] parameters. The joint-based regressor's graph convolutional layers refines the image features of joint predictions by fully exploiting the human kinematic prior and regress parameters of a 3D mesh that best describes a target person in a crowd. The fourth and fifth rows of Figure A prove that our approach is also effective on estimating robust 3D meshes from truncated images, which often have missing 2D joint predictions.

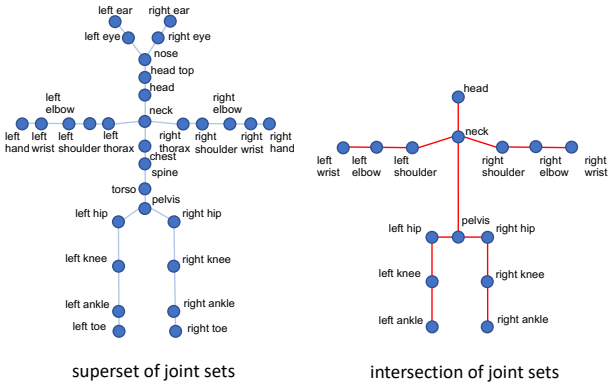

Figure B. Illustration of joint sets. The red skeleton of the intersection of joint sets defines the joints' neighborhood of graph convolution used in the joint-based regressor.

Comparison with SPIN. We provide more qualitative comparison with SPIN [11] in Figure C. SPIN is one of the state-of-the-art methods that are based on the two wheels of the current 3D human mesh estimation literature, the mixed batch training and the model-based approach using a global feature discussed in Section 1 of the main manuscript. Our 3DCrowdNet produces accurate and robust 3D meshes from diverse in-the-wild crowded scenes. On the other hand, SPIN predicts an incorrect overall pose for a person under severe inter-person occlusion (top-left), estimates inaccurate leg poses (bottom-left, bottom-right), and produces

Figure C. Qualitative comparison on the CrowdPose [12] test set. We highlighted the failure cases of SPIN [11] with red circles. SPIN tends to be sensitive to occlusion, while 3DCrowdNet provides robust 3D meshes.

Figure D. Qualitative comparison on the CrowdPose [12] test set. We highlighted the failure cases of ROMP [19] with red circles. Wrong global rotation of occluded persons (the third and fourth rows); inaccurate leg poses under inter-person occlusion (the first and third rows). 3DCrowdNet produces much more robust 3D meshes.

noisy 3D meshes (top-right, bottom-left, bottom-right) that show vulnerability to inter-person occlusion.

Comparison with ROMP. We provide more qualitative comparison with ROMP [19] in Figure D. ROMP is a bottom-up method for multi-person 3D mesh estimation. Our 3DCrowdNet produces accurate and robust 3D meshes from diverse in-the-wild crowded scenes. On the contrary, ROMP predicts the wrong global rotation of a target per-

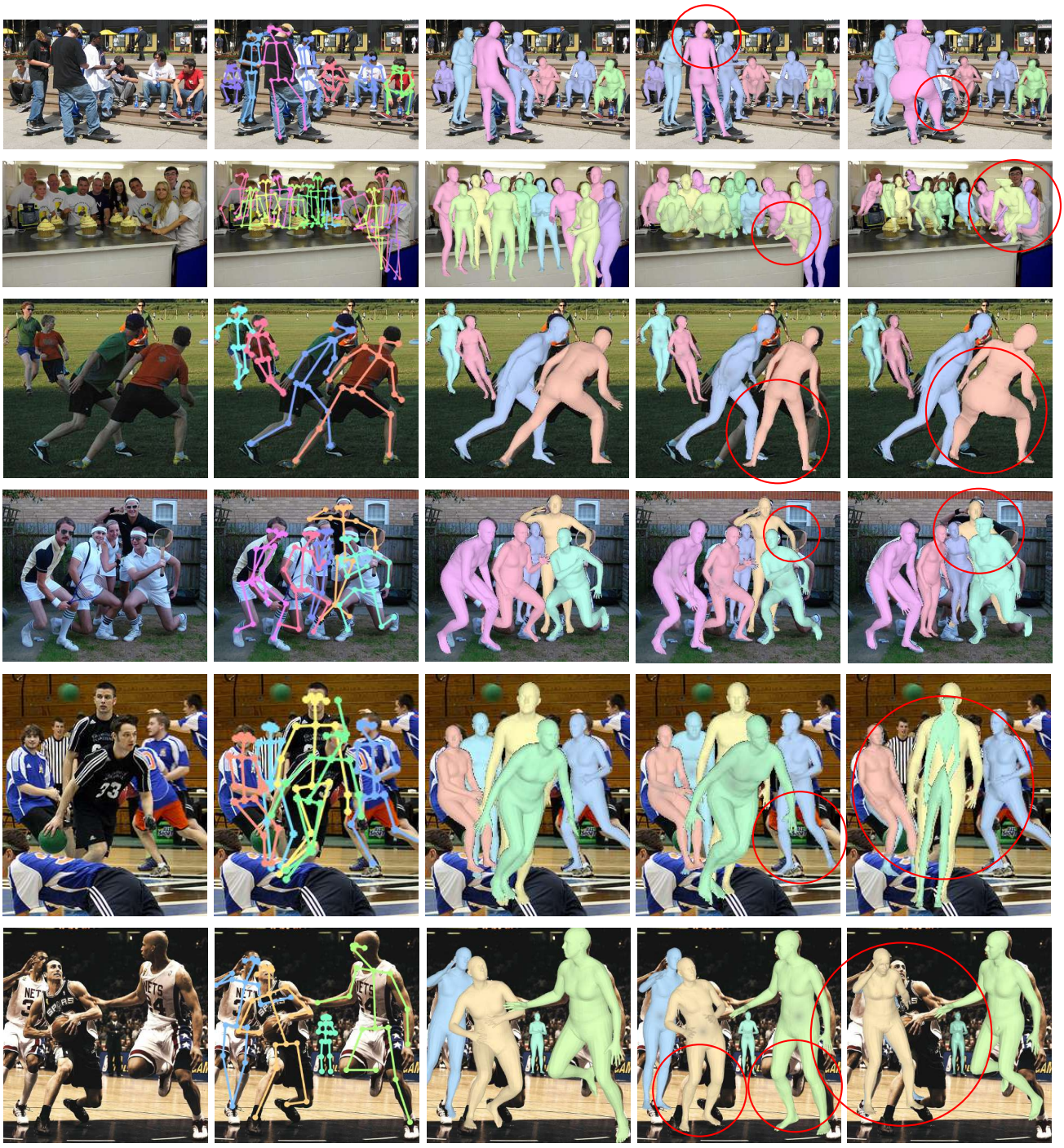

Figure E. Qualitative comparison on the CrowdPose [12] test set. From left, an input image, 2D pose input, 3DCrowdNet, I2L-MeshNet [17], and Pose2Mesh [5] outputs. Our 3DCrowdNet successfully disentangles a target person from other people in a bounding box compared with I2L-MeshNet. Also, 3DCrowdNet produces a 3D shape that best describes a target person in images, while Pose2Mesh estimates a plausible 3D shape for given 2D poses, which does not correspond to input images. 3DCrowdNet and I2L-MeshNet use the same bounding boxes to crop an image for each person. 3DCrowdNet and Pose2Mesh use the same 2D poses from [4].

Figure F. Failure cases of 3DCrowdNet.

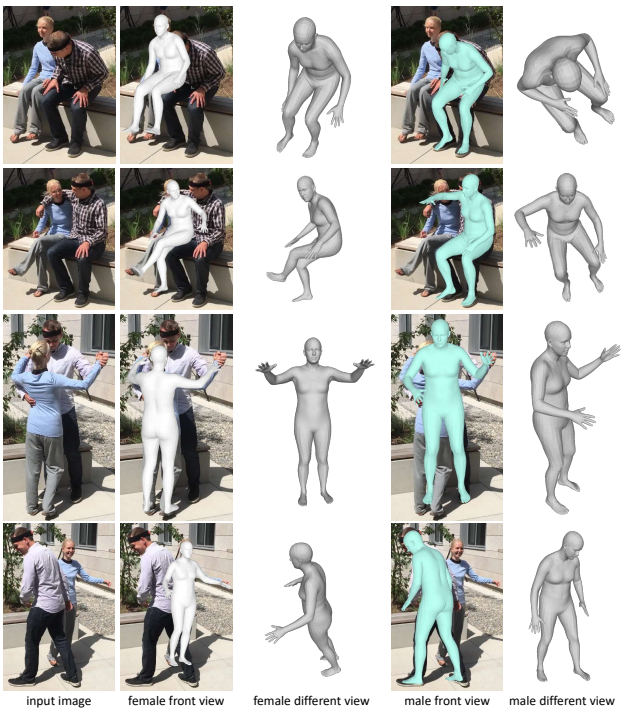

Figure G. 3DCrowdNet's outputs on 3DPW-Crowd.

son from in-the-wild crowded scenes (the third and fourth rows) and produces inaccurate leg poses under severe inter-person occlusion and crossed human parts (the first and second rows).

Comparison with Pose2Mesh and I2L-MeshNet. Figure E shows the qualitative comparison between 3DCrowdNet, Pose2Mesh [5], and I2L-MeshNet [17]. Pose2Mesh and I2L-MeshNet are state-of-the-art model-free 3D mesh estimators, which predict coordinates of mesh vertices. Especially, Pose2Mesh is one of the most relevant competitors, since it can also benefit from the same 2D pose input. 3DCrowdNet produces much more robust 3D meshes from in-the-wild crowded scenes than the two methods. Pose2Mesh estimates the most plausible 3D mesh for a given 2D pose (the first and fourth rows), not the 3D mesh that best describes a target in a crowd, as discussed in Section 5 of the main manuscript. Also, it often wrongly corrects the 2D pose input and produces common standing leg poses different from the images (the third, fifth, and sixth rows). I2L-MeshNet fails to distinguish different people in overlapping bounding boxes (the fifth and sixth rows). In addition, it tends to provide very noisy 3D pose and shape of a target in crowded scenes, which reveals the method's vulnerability to inter-person occlusion. The results in the second row also validate the superiority of 3DCrowdNet's robustness to truncated bodies.

Results on 3DPW-Crowd. Figure G illustrates the 3DCrowdNet's outputs on 3DPW-Crowd. 3DCrowdNet estimates robust 3D pose and shape on images that show people having highly close interaction. Different people in overlapping bounding boxes are disentangled, and occluded body parts are reasonably reconstructed.

C. Limitation

Although the proposed 3DCrowdNet highly outperforms the previous 3D mesh estimation methods in in-the-wild crowded scenes, there is a limitation to be resolved in future work. As shown in Figure F, when the 2D pose is inaccurate and appearances of nearby persons are very similar, 3DCrowdNet fails to produce robust 3D meshes. The top-left and bottom-right cases of Figure F are the representative cases, which can be easily found in sports images. In

such cases, it is challenging for 3DCrowdNet to correct the inaccurate 2D pose with image features, since the context information in image features is ambiguous due to indistinguishable appearances. One way of resolving the challenge could be to model the relative translation between persons to better understand the context. Alternatively, data augmentation to make a network robust to similar appearances would be an interesting direction.

D. 2D pose estimators.

In this work, we used 2D pose outputs from OpenPose [3] and HigherHRNet [4]. The OpenPose outputs used in 3DPW [20] are included in the annotations of 3DPW [20]. The OpenPose used in MuPoTS [16] are obtained by running the third-party PyTorch [18] code implementation[1]. OpenPose is trained on **COCO2017 train** [13] dataset. It achieves 65.3 mAP (mean Average Precision) in **COCO2017 val** dataset. In the CrowdPose [12] test set, it achieves 48.7 and 32.3 mAPs for medium and hard cases, respectively. All the HigherHRNet outputs are obtained by running the official code implementation. HigherHRNet is trained on **COCO2017 train** dataset. It achieves 0.671 mAP on **COCO2017 val** dataset. In the CrowdPose [12] test set, it achieves 68.1 and 58.9 mAPs for medium and hard cases, respectively.

License of the Used Assets

- MSCOCO dataset [13] belongs to the COCO Consortium and are licensed under a Creative Commons Attribution 4.0 License.
- Human3.6M dataset [6]'s licenses are limited to academic use only.
- MPII dataset [2] is released for academic research only and it is free to researchers from educational or research institutes for non-commercial purposes.
- 3DPW dataset [20] is released for academic research only and it is free to researchers from educational or research institutes for non-commercial purposes.
- CrowdPose dataset [12] is released for academic research only and it is free to researchers from educational or research institutes for non-commercial purposes.
- MuCo-3DHP and MuPoTS [16] are released for any non-commercial purposes.
- CMU-Panoptic [10] is released only for research purposes.
- The third party implementation[2] of OpenPose [3] is licensed under the MIT license.
- HigherHRNet [4]'s implementation[3] is licensed under the MIT license.

[1] https://github.com/tensorboy/pytorch_Realtime_Multi-Person_Pose_Estimation

[2] https://github.com/tensorboy/pytorch_Realtime_Multi-Person_Pose_Estimation

[3] https://github.com/HRNet/HigherHRNet-Human-Pose-Estimation

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 1

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1, 6

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 2, 6

[4] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 1, 2, 4, 6

[5] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, 2020. 1, 2, 4, 5

[6] 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. 1, 2, 6

[7] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 1, 2

[8] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 1

[9] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 1

[10] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic Studio: A massively multiview system for social interaction capture. *TPAMI*, 2017. 6

[11] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1, 2, 3

[12] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019. 1, 2, 3, 4, 6

[13] 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. 1, 6

[14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 2

[15] 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. 1

[16] 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. 1, 6

[17] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-Lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*, 2020. 1, 2, 4, 5

[18] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NeurIPS*, 2017. 6

[19] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *ICCV*, 2021. 1, 3

[20] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 1, 2, 6

[21] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. AI Challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 1

[22] 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. 1

[23] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 1