Supplementary Material *for* JOTR : 3D Joint Contrastive Learning with Transformers for Occluded Human Mesh Recovery

Jiahao Li[1,2*], Zongxin Yang[1], Xiaohan Wang[1], Jianxin Ma[2], Chang Zhou[2], Yi Yang[1]
[1] ReLER, CCAI, Zhejiang University [2] DAMO Academy, Alibaba Group

Method	Conference	Params
SPIN [15]	ICCV 2019	27.0M
PyMaf [27]	ICCV 2021	45.2M
PARE [14]	ICCV 2021	32.9M
ROMP [21]	ICCV 2021	33.1M
3dCrowdNet [6]	CVPR 2022	30.5M
OCHMR [13]	CVPR 2022	35.8M
Ours	-	39.6M

Table A: Comparisons to the state-of-the-art methods on model parameters. The amount of model parameters is comparable with previous methods. However, our proposed method recovers meshes by fusing 2D and 3D features obtained from lifting 2D features to 3D. To learn and reason 3D representations, our JOTR incorporates an additional module, which increases the number of parameters.

In this supplementary material, we provide additional implementation details, ablation studies and qualitative results in Section A and B and C respectively.

A. Additional Implementation Details

A.1. Training Details

Implementation Details. In our experiments, we apply the pretrained weights by Xiao *et al*. [26] to initialize ResNet-50 [7] because of the slow convergence with ImageNet pretrained weights as analyzed by [6, 14]. As for transformers, the channel size of the input and output is 256, and the number of heads is 8. The number of layers is 2, 2, 1 and 3 for 2D transformer encoder, 3D transformer encoder, transformer decoder and refining layers respectively, and the dimension of the feed-forward network is 1024. SMPL query tokens and joint query tokens are randomly initialized and updated during training. As for positional encoding, we apply the sine-cosine positional encoding [4, 23] to the input of the 2D and 3D transformer encoders. For training speedup, we apply distributed training with PyTorch [20] using 8 Tesla V100 GPUs.

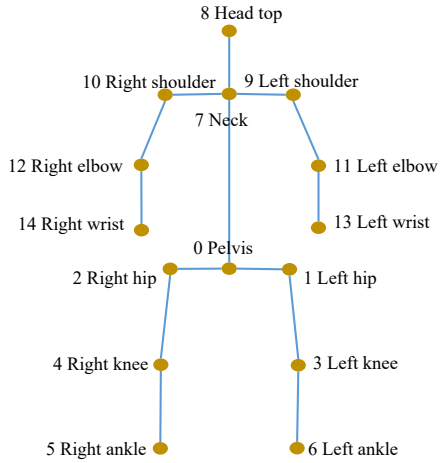

Figure A: Human body joints we use in our experiments for locating and extracting 3D joint features in 3D space.

Loss. We apply trainable parameters to $\lambda_{3D}, \lambda_{2D}$ and λ_{SMPL}, and the parameters are initialized as 1.0. λ_{j2n} and λ_{j2j} are set to 0.01.

Human Body Joints.

As shown in Fig. A, we employ 15 joints to locate and extract 3D human representation in 3D space. Note, the joints are not applied for evaluating the performance of our method, but for adding supervisions for 3D space.

A.2. Details of Evaluating Datasets

3DPW-OC. Following Zhang *et al*. [29], 3DPW-OC contains 23 person-object occlusion video sequences. Please refer to [29] for the details.

3DPW-PC. 3DPW-PC contains 1314 frames of 6 person-person occlusion video sequences. They contain severe person-person occlusion cases with at least two-people overlapping. Please refer to [21] for the details.

3DPW-Crowd. 3DPW-Crowd contains 1073 frames of 2

Method	Training Datasets	MPJPE ↓
ROMP [21]	Human3.6M [8], MPI-INF-3DHP [18], MuCo-3DHP [19], MSCOCO [17], CrowdPose [16], 3DOH [29], MPII [2], LSP [10], LSP-Extended [11], AICH [25], PoseTrack [1]	134.6
3dCrowdNet [6]	Human3.6M [8], MuCo-3DHP [19], MSCOCO [17], CrowdPose [16], MPII [2]	127.6
BEV [22]	Human3.6M [8], MuCo-3DHP [19], MSCOCO [17], CrowdPose [16], MPII [2], LSP [10]	127.9
Ours	Human3.6M [8], MuCo-3DHP [19], MSCOCO [17], CrowdPose [16]	**114.7**

Table B: Comparisons to the state-of-the-art methods on CMU-Panoptic [12]. Our proposed JOTR uses the least training datasets and achieves the best accuracy in multi-person crowded scenes.

Figure B: An example of internet video. **Please use Adobe Acrobat to view it.**

Figure C: An example of internet video. **Please use Adobe Acrobat to view it.**

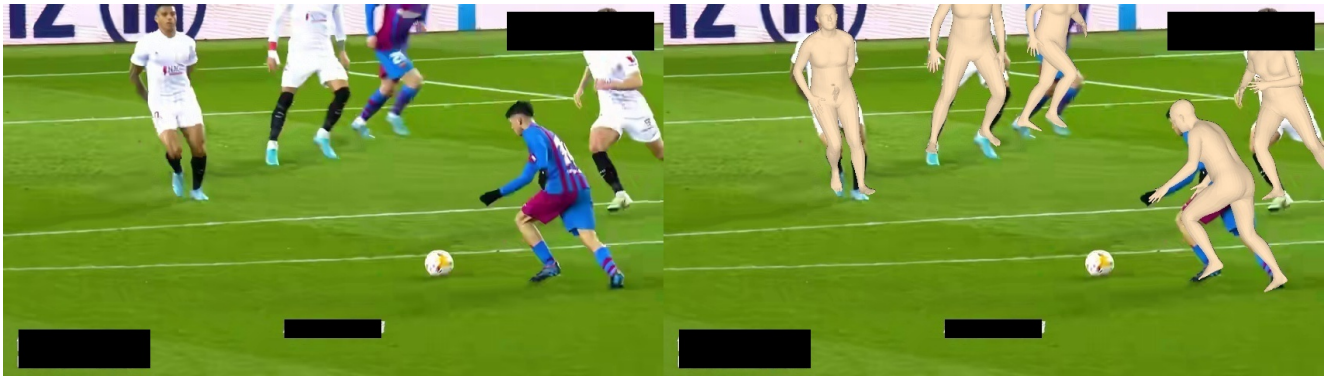

Figure D: An example of internet video. **Please use Adobe Acrobat to view it.**

person-person occlusion video sequences. Please refer to [6] for the details.

3DOH. 3DOH [29] is an object-occluded dataset containing 1290 images for testing, which is used to evaluate the performance under object occlusion.

CMU-Panoptic [12] is a dataset with multi-person indoor

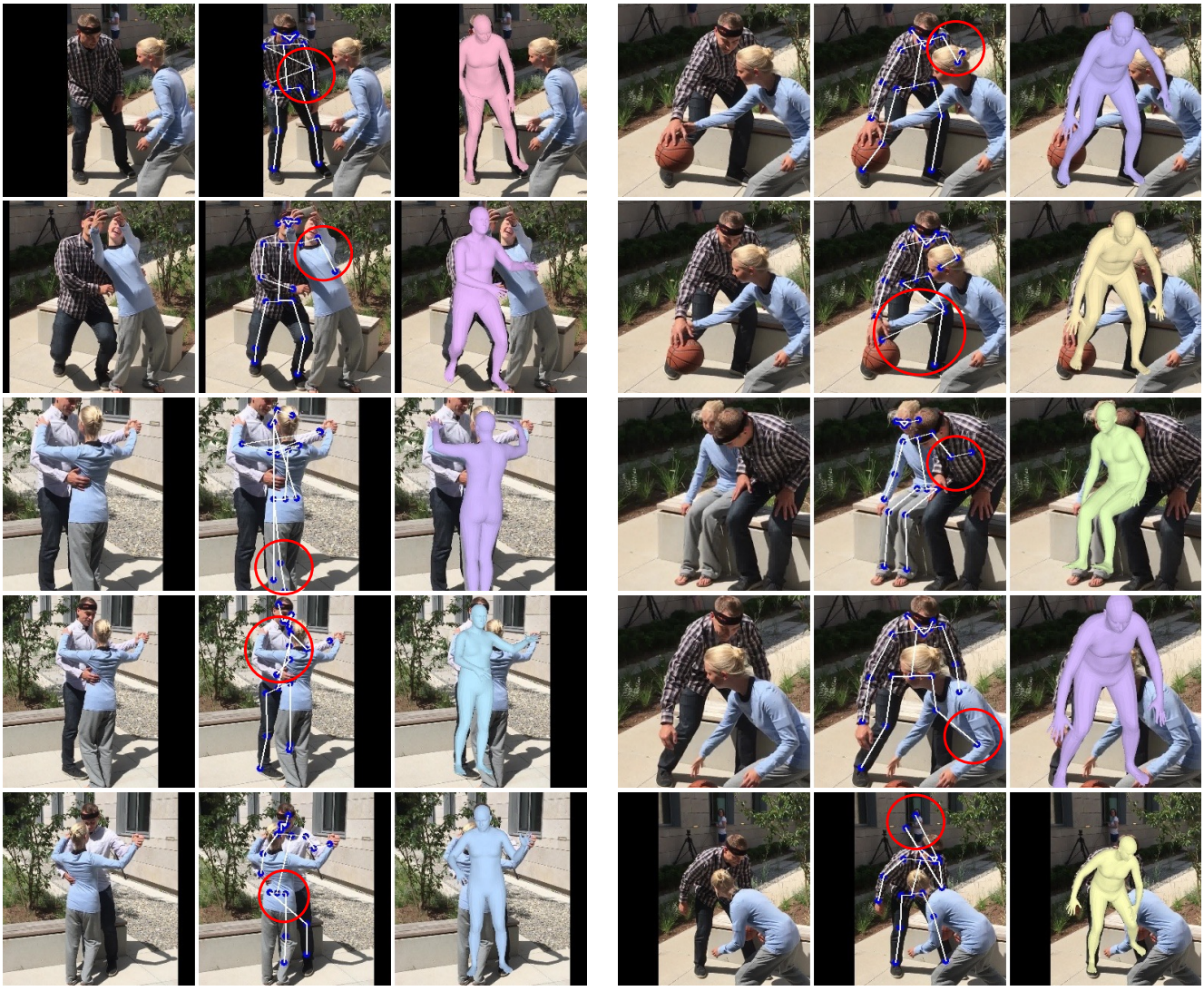

Figure E: Visualization of our predictions under incorrect 2D poses. Our JOTR can still recover the correct 3D pose and shape.

J2N	J2J	MPJPE ↓	PA-MPJPE↓	PVE↓
✗	✗	84.7	53.2	106.1
✓	✗	82.8	52.6	104.2
✗	✓	83.2	52.2	104.6
✓	✓	**82.4**	**52.0**	**103.4**

Table C: Ablation study of 3D joint contrastive learning on 3DPW-Crowd. J2N: joint-to-non-joint contrast. J2J: joint-to-joint contrast.

Positive Samples	MPJPE ↓	PA-MPJPE↓	PVE↓
Predicted Joints	83.6	52.5	104.9
GT Joints	83.2	52.2	104.6

Table D: Ablation study of joint-to-joint contrastive learning on 3DPW-Crowd.

scenes. Following [9, 21], we select four sequences (*i.e.*, *Haggling*, *Mafia*, *Ultimatum*, and *Pizza*) for evaluation. The sequences contain 9600 frames and 21,404 persons with GT 3D pose annotations.

3DPW [24] We use the test set of 3DPW [24] following the official split protocol. The test set contains 26240 images and 35515 persons with GT 3D pose and shape annotations. We use 14 joints defined by Human3.6M [8] for evaluating PA-MPJPE and MPJPE following the previous works [9, 21, 29].

Figure F: Qualitative comparison on the OCHuman [28]. 3DCrowdNet and PARE fails to recover the details of human body due to the lack of estimating invisible information. Our JOTR mitigates this limitation by reasoning in both 2D and 3D features.

A.3. Details of 2D Pose Predictors

To guide our 3D mesh reconstruction, we utilize 2D pose outputs from OpenPose [3] and HigherHRNet [5]. OpenPose outputs are used for 3DPW, 3DPW-OC, and 3DPW-PC since they are included in the annotations. As for 3DPW-Crowd, 3DOH and CMU-Panoptic, we use the HigherHRNet outputs by running the official code implementation. Note, for a fair comparison, we use the same 2D pose input for both 3DCrowdNet and our JOTR.

B. Additional Ablation Studies

All experiments of ablation studies are carried out on the 3DPW-OC dataset, as described in the main manuscript (*i.e.*, Tab. 4 5 6 and 7).

3D joint contrastive learning on 3DPW-Crowd. As shown in Tab. C, we conduct ablation studies of 3D joint contrastive learning on person-person occlusion scenarios. Both joint-to-non-joint and joint-to-joint contrastive losses result in improved performance, which indicates that our proposed 3D joint contrastive learning also works well in person-person occlusion scenarios.

Positive samples in joint-to-joint contrast. A natural question is that whether the positive samples should be the predicted joints or the ground-truth joints. As shown in Tab. D, we conduct an ablation study on this question. We find that using the predicted joints as positive samples results in a slight performance drop. A possible reason is that the predicted joints are not accurate enough in the early training stage, resulting in extracting less informative joint features, which leads to the performance drop.

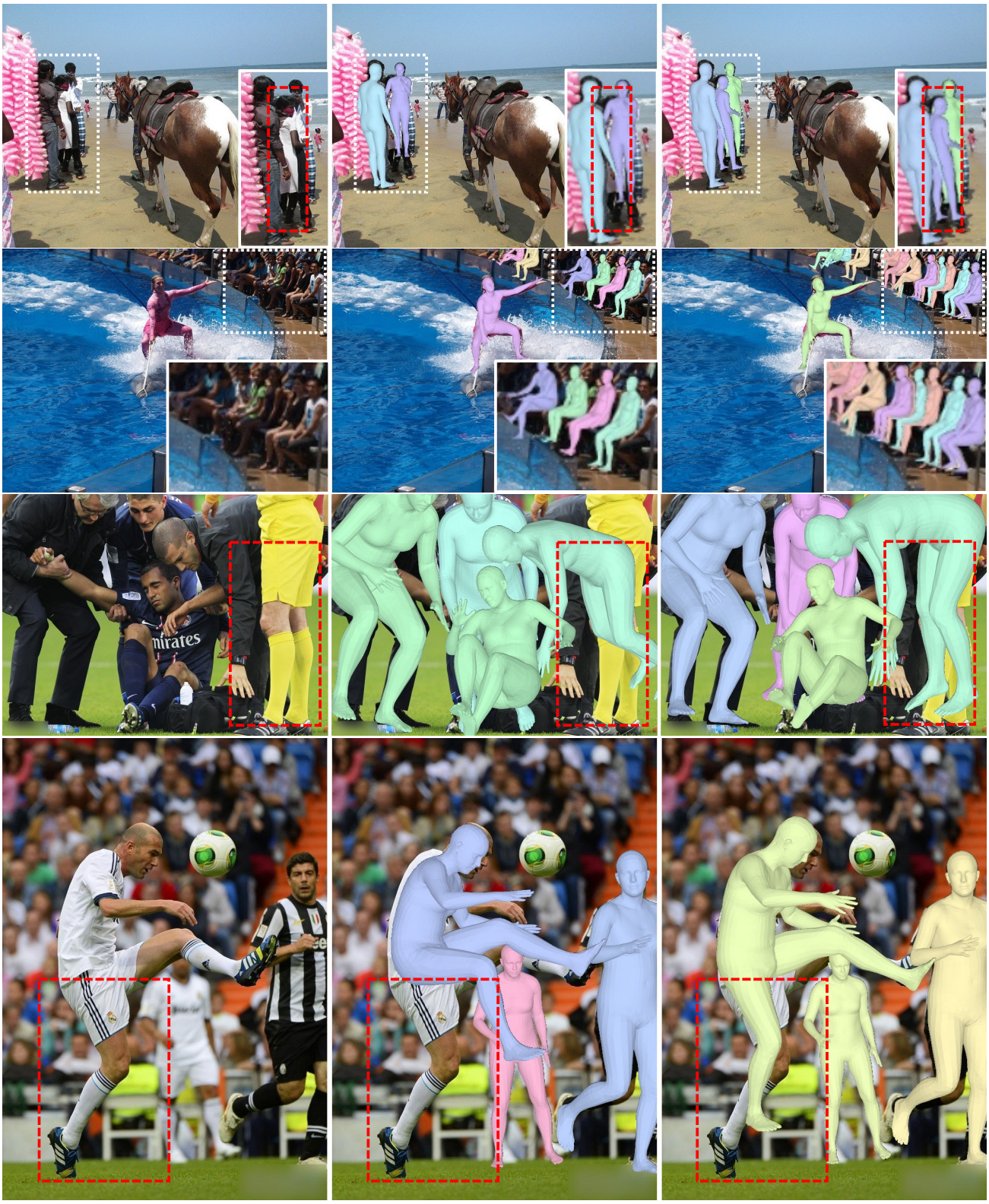

Figure G: Qualitative comparison on the CrowdPose [16] test set.

Input Image ROMP JOTR (Ours)

Figure H: Qualitative comparison on the CrowdPose [16] test set.

C. Additional Qualitative Results

Qualitative results on challenging Internet videos. The qualitative results obtained from challenging Internet videos are illustrated in Fig. B, C and D. Due to the limitation of the detector, some frames remain persons not detected. Notably, our JOTR method achieves commendable performance without any employing temporal smoothing techniques.

Accurate 3D predictions from inaccurate 2D inputs. As shown in Fig. E, in the case of person-person occlusion, the 2D pose outputs from 2D pose detectors may be inac-

curate. However, our JOTR can still recover accurate 3D human meshes from these inaccurate 2D pose inputs, which demonstrates the robustness of our JOTR.

Comparison with 3DCrowdNet and PARE. 3dCrowdNet [6] is a top-down mesh recovery method, which also applies 2D keypoints as guidance. As shown in Fig. F and G, even though 3dCrowdNet adds 2D keypoints in the input, it still fails to recover the 3D human mesh in the case of severe occlusion due to the lack of 3D information. Our JOTR not only recovers accurate 3D human meshes under occlusions, but also predicts the possible poses of unseen body parts.

PARE [14] is a top-down mesh recovery method for person-object occlusion scenarios. As shown in Fig. F, it fails under multi-person crowding scenarios because of the noise from other people. Also, it suffers from the lack of estimating invisible information resulting in unreasonable predictions of unseen body parts.

Comparison with ROMP. ROMP [21] is a bottom-up mesh recovery method for multi-person scenarios. Fig. H shows qualitative comparison with ROMP. ROMP samples 2D features via center points to recover 3D meshes, which is not robust enough to describe the detail of the human body and sensitive to occlusions. Our method fuses 2D and 3D features with transformers and attends 3D joints behind occlusions, resulting in robust 3D mesh recovery.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 2

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 4

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[5] 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. 4

[6] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1475–1484, 2022. 1, 2, 7

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

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

[9] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020. 3

[10] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Aberystwyth, UK, 2010. 2

[11] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472. IEEE, 2011. 2

[12] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 2

[13] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1715–1725, 2022. 1

[14] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 1, 7

[15] 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

[16] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 2, 5, 6

[17] 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. 2

[18] 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 *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 2

[19] 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 *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, sep 2018. 2

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[21] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021. 1, 2, 3, 7

[22] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 2

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 1

[24] 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 *European Conference on Computer Vision (ECCV)*, sep 2018. 3

[25] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1480–1485. IEEE, 2019. 2

[26] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1

[27] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11446–11456, 2021. 1

[28] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2019. 4

[29] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7376–7385, 2020. 1, 2, 3