Supplementary Material for "Coherent Reconstruction of Multiple Humans from a Single Image"

Wen Jiang[2*], Nikos Kolotouros[1*], Georgios Pavlakos[1], Xiaowei Zhou[2†], Kostas Daniilidis[1]
[1] University of Pennsylvania [2] Zhejiang University

This supplementary material includes additional details that were not included in the main manuscript due to space constraints. We start with more implementation details (Section 1). We continue with a short discussion about the effect of the model and the losses in our approach (Section 2). Then, we provide further results from our quantitative experiments (Section 3). Finally we extend our qualitative evaluation, including more examples of our approach, including successes, failures and comparisons with the baseline model (Section 4).

1. Implementation details

1.1. Architecture

Our architecture follows the typical Faster-RCNN pipeline [24], where we add an additional branch for SMPL parameter regression. This branch follows the architecture choices of the iterative regressor proposed by Kanazawa *et al*. [10]. Ultimately, the output of the SMPL branch includes the estimated pose and shape parameters for the corresponding bounding box, θ and β respectively, as well as the camera parameters $\boldsymbol{\pi} = \{s, x, y\}$. In the original HMR formulation [10], the camera parameters include a scaling factor s, as well as a 2D translation t_x, t_y for a weak perspective camera. However, in order to produce a coherent scene we need to move away from the original weak perspective camera assumption. To do that, we propose a way of converting the camera parameters $\boldsymbol{\pi}$ to the actual translation of each person in the scene.

Let us represent with M_i, $\boldsymbol{\pi}_i = \{s_i, x_i, y_i\}$, the regressed mesh and camera parameters respectively for the ith bounding box B_i in an image I with width w and height h. For each image, we assume we have a single camera located at the origin of the coordinate system with focal length f and its principal point at the center of the image. We underline that the camera parameters we regress are not for weak perspective projection. Instead, we assume a fully perspective camera model, where the focal length f is known. Let $B_i = [x_{min}, y_{min}, x_{max}, y_{max}]$, with center $c_i = [(x_{min} + x_{max})/2, (y_{min} + y_{max})/2]$ and size $\alpha_i = \max(x_{max} - x_{min}, y_{max} - y_{min})$. Given these parameters, the depth of the person is calculated as:

$$d_i = \frac{2f}{s_i \alpha_i} \tag{1}$$

Using the computed depth, we then define the person translation as:

$$t_i = \begin{bmatrix} d_i \left(x_i \alpha_i + c_{i,x} - w/2\right)/f \\ d_i \left(y_i \alpha_i + c_{i,y} - h/2\right)/f \\ d_i \end{bmatrix} \tag{2}$$

The above transformation performs a "coordinate change" from the local, per-bounding box camera to the single global scene camera. This choice ensures that the projection of $\hat{t}_i = [x_i, y_i, d_i]$ given a camera with principal point at the center of the bounding box, projects to the same point, as t_i given a camera with principal point in the image center. Intuitively, the SMPL branch predicts camera parameters for each box independently. These parameters are relative to the bounding box size, because the input to the SMPL head is the 14×14 output of the ROI Align, so they have to be scaled accordingly.

1.2. Interpenetration loss

Here we will elaborate more on how the interpenetration loss works in cases where there are collisions between different people. In the main text we defined the interpenetration loss for a scene as:

$$L_{\mathcal{P}} = \sum_{j=1}^{N} \rho \left(\sum_{i=1, i \neq j}^{N} \mathcal{P}_{ij} \right). \tag{3}$$

As explained in the main text, the loss for each person is applied at the vertex level; for person j, we penalize all the vertices that lie inside another person i and that penalty specifically is:

$$\mathcal{P}_{ij} = \sum_{v \in M_j} \tilde{\phi}_i(v). \tag{4}$$

* Equal contribution.
† X. Zhou and W. Jiang are affiliated with the State Key Lab of CAD&CG, Zhejiang University. Email: xwzhou@zju.edu.cn.

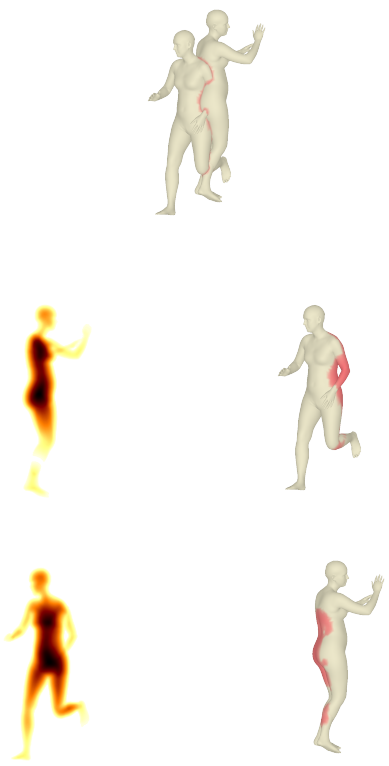


Figure 1: **Illustration of interpenetration loss.** Top: Collision between two people. Center: Distance field ϕ_2 for person 2 and penalized vertices of person 1. Bottom: Distance field ϕ_1 for person 1 and penalized vertices of person 2.

Because the loss is applied at the vertex level, $\mathcal{P}_{ij}$ is not symmetric. This is depicted with an example showing a collision between two people in Figure 1.

1.3. Training strategy

We observed that the tasks of detection and 3D shape reconstruction behave quite differently during training, with the reconstruction branch needing significantly more training iterations than the detection branch. For this reason, before training the full network end-to-end, we pretrained the SMPL head with cropped images for roughly 350K iterations. The pretraining was done using single-person examples from Human3.6M [5], MPI-INF-3DHP [18], COCO [13], LSP [6], LSP Extended [7] and MPII [1]. After this step, training continues with multi-person images for 400k more iterations. For our full model, our proposed losses are also active in this second step, while for the results reported as "baseline", they are not. We trained our full model in 2 1080Ti GPUs with a learning rate of $1e-4$ using the Rectified Adam optimizer [14]. Regarding the weights for the different loss functions, we use 4 for the keypoint reprojection loss, 4 for the 3D keypoint loss, 1 for the loss on the SMPL θ parameters, 1/100 for the loss on the SMPL β parameters, 1/60 for the adversarial prior, 1/100 for the collision loss, 100 for the depth ordering loss, and 1 for the detection and RPN losses.

Before applying the 2D keypoint loss, we normalize the keypoints inside each box proposal by subtracting the box center and dividing by the box width. The other losses of the SMPL branch (3D keypoint loss, loss on the SMPL parameters, loss on the adversarial prior) are computed the same way as in HMR [10].

2. Effect of model and losses

For the ResNet50 backbone, each neuron has a receptive field of size 483×483 pixels. This means that for a $h \times w$ bounding box, the receptive field is $(h+482) \times (w+482)$. Considering that the input images have resolution 512×832 pixels, for a bounding box, we expect most of its neighboring people to be within its receptive field. The most interesting scenario is when we have three people, A, B and C, where A overlaps with B, B overlaps with C, but A does not overlap with C. This is challenging, since C might not be "visible" from A and vice versa. In that case, it would be hard to get a coherent prediction for the whole group. To examine how often this occurs, we investigated the statistics of the datasets we used. Specifically we focused in this (A,B,C), scenario, where B can correspond to more than one person, i.e., we can have a longer chain. Considering the receptive field of our architecture, we observed that in most cases C is "visible" from A. Particularly, across all the cases where this (A,B,C) scenario happens, in 88% of cases for Panoptic, in 92% of cases for MuPoTS, and in 91% of cases for PoseTrack, person C is included in the receptive field of A. We expect that with a deeper network, that has a larger receptive field, e.g., ResNet152, these percentages will be even higher.

Regarding our proposed losses, they belong in the category of cross-instance supervision. Cross-instance losses have also been applied successfully in recent works, e.g., [4, 12]. Effectively, during training, to decrease these losses, the network needs to develop features that help avoid coherency errors. The learned features can be related to depth, occlusion, segmentation of the person, etc. Since the losses decrease during training, the network does generate helpful features. More importantly, this translates also to improvement at test time (Tables 4,5 in main manuscript present improvements in *unseen* datasets particularly for collisions and depth ordering). This is a strong indication that the network is not overfitting, but it is indeed learning features that generalize across scenes, and encourage it to make coherent predictions *at test time too*.

Method	All	Matched
Our baseline	66.95	68.96
Our baseline + $L_{\mathcal{P}}$	67.84	70.00
Our baseline + $L_{\mathcal{D}}$	66.59	68.43
Our baseline + $L_{\mathcal{P}}$ + $L_{\mathcal{D}}$	**69.12**	**72.22**

Table 1: **Ablative on MuPoTS-3D**. The numbers are 3DPCK. We report the overall accuracy (All), and the accuracy only for person annotations matched to a prediction (Matched).

3. Quantitative results

First we provide a more detailed evaluation of our proposed losses on the MuPoTS-3D dataset [19]. We have already reported the results of our baseline and our full model in Table 3 of the main manuscript, but here we extend to a more fine-grained ablative study. The complete results for different versions of our model are presented in Table 1. Based on the results, we see that the use of the interpenetration loss alone improves slightly the results over the baseline, while with the depth ordering-aware loss alone we observe a small decrease in the accuracy. However, when we combine the two losses together, we achieve better results, both compared to the baseline, as well as compared to the versions using only one of the two losses alone.

Regarding the comparison with the state-of-the-art in the main manuscript, our evaluation has focused on approaches that estimate 3D pose and shape in the form of the SMPL parametric model [16]. This is common in the literature, where SMPL-based approaches, e.g., [10, 21, 23] do not directly compare with skeleton-based approaches, e.g. [17, 27], and vice versa, because of the different output they provide. Typically, skeleton-based approaches report better quantitative results when compared on metrics using 3D joints, but SMPL-based approaches still output a more informative representation in the form of 3D rotations for each part, making the task harder than only estimating 3D joint locations. Although we are not directly comparable with skeleton-based approaches, we observe that on MuPoTS-3D [19] our approach still performs better than [25, 19], it is competitive to [26] and is underperforming only when it is compared to the most recent baseline, i.e., [20]. However, this comparison is done under the traditional 3D pose metrics, which are computed on 3D joints of individual people only. When the evaluation is performed on a metric that requires coherent estimation for all the people in the scene, e.g., on depth ordering, we observed that even the state-of-the-art approach of Moon *et al*. [20] performs worse than our approach. Concerning the evaluation with the single-person pose and shape baselines our comparison focuses primarily on HMR [10], that is more similar to us in terms of architecture, training details, and training data. Some more recent approaches, e.g., [3, 11, 22, 28] report improved results on single-person datasets, but they rely on improved training techniques or architectures. These improvements are orthogonal to ours, since we focus on improving the multi-person results, and not the single-person case as they do.

For the evaluation on MuPoTS-3D, we only presented the mean accuracy over all sequences. Here we also provide a more detailed evaluation for each sequence. The complete results are included in Table 2. As we can see, for most sequences, and overall, the version of our model trained with the proposed losses outperforms our baseline.

Besides the above experiments, we also present additional ablatives to clarify the effect of using different datasets to train our system. Similar to Kanazawa *et al*. [10], we use a large set of datasets to train our network, since we observed that this diverse set of images is helpful for better generalization to in-the-wild settings. However, for simpler indoor settings, like Human3.6M [5] and Panoptic [8, 9], using only COCO [13] and Human3.6M [5] for training provides comparable results. To focus on the effect of the data specifically on pose reconstruction, we investigate a simpler setting where we train only the ResNet backbone and the SMPL head, providing ground truth bounding boxes during testing. As we can see in Table 3 for Panoptic and Table 4 for Human3.6M, for these indoor datasets, training with all the data achieves similar performance with training only with Human3.6M and COCO. It is also interesting to observe that using ground truth bounding boxes instead of detections improves performance for Human3.6M, but it hurts performance on Panoptic. This can be attributed to the fact that Panoptic has many truncated human instances, so learning to jointly crop the most informative bounding box along with reconstructing the person can be beneficial compared to be given an arbitrary bounding box at test time.

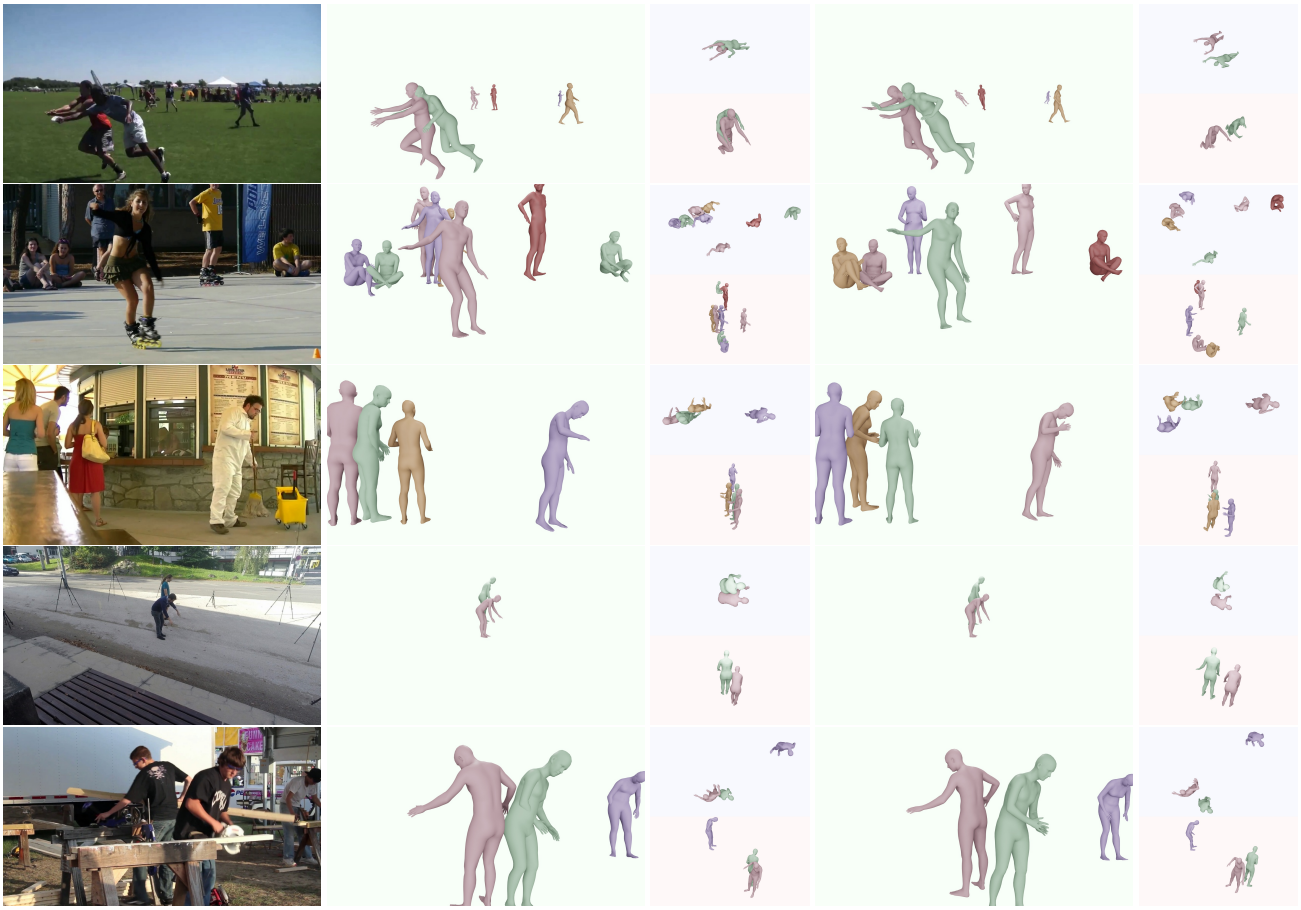
Additionally, we also ablate the type of supervision we use for Human3.6M. Similar to [10], we use SMPL parameters provided by fitting SMPL to surface markers through MoSh [15]. To see if we can relax this constraint, we also use SMPL parameters provided by fitting SMPL to Human3.6M 3D keypoints, using a procedure similar to SMPLify [2]. Again the results are comparable (Tables 3 and 4), which means that our performance does not rely explicitly on the availability of MoSh parameters.

4. Qualitative results

For our qualitative evaluation, in Figure 2, we provide more comparisons between our baseline model and our full model trained with our proposed losses. Then, in Figures 3 and 4 we provide more successful reconstructions from the datasets we use in our evaluation. Finally, in Figure 5 we present some representative failure cases of our approach.

	Method	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS18	TS19	TS20	Avg
All	Ours (baseline)	76.42	65.75	**71.59**	66.26	76.89	32.89	**74.01**	67.68	**60.52**	**78.88**	57.81	55.55	64.38	59.68	70.87	75.01	69.84	69.60	**75.19**	70.18	66.95
	Ours (full)	**80.60**	**68.65**	67.02	**68.19**	**77.78**	**38.99**	**74.01**	**67.88**	54.69	77.11	**63.77**	**64.73**	**64.40**	**60.37**	**72.71**	**83.68**	**75.53**	**76.91**	74.40	**70.67**	**69.12**
Matched	Ours (baseline)	76.42	71.91	**71.77**	66.48	79.16	32.92	74.30	68.93	**60.52**	**78.88**	57.81	55.55	66.80	70.80	70.87	75.71	69.95	73.08	76.55	80.89	68.96
	Ours (full)	**80.60**	**76.59**	67.19	**68.42**	**80.24**	**40.33**	**74.71**	**70.77**	54.69	77.11	**64.88**	**64.73**	**67.92**	**72.91**	**72.84**	**85.03**	**75.97**	**81.89**	**78.52**	**89.04**	**72.22**

Table 2: **Full results on MuPoTS-3D**. The numbers are 3DPCK. We report the overall accuracy (All), and the accuracy only for person annotations matched to a prediction (Matched).



Input image Baseline Ours

Figure 2: **Qualitative effect of our proposed losses.** Given an input image (first column), we provide results of the baseline model (second and third column) and our full model trained with our proposed losses (fourth and fifth column). As expected, we improve over our baseline in terms of coherency in the results (i.e., fewer interpenetrations, more consistent depth ordering for the reconstructed meshes). For the first image, the visualization focuses only on the two people in the foreground and the rest are ignored.

Data	MoSh	Haggling	Mafia	Ultim.	Pizza	Mean
All data	Yes	155.4	178.6	179.7	186.1	175.0
COCO+H36M	Yes	157.5	180.3	178.3	191.7	177.0
COCO+H36M	No	158.7	176.4	175.0	190.4	175.1

Table 3: **Ablative on the Panoptic dataset**. We focus on the ResNet backbone and the SMPL head (i.e., we use ground truth bounding boxes) and we ablate different training strategies; using all training data (first row), reducing the training data to COCO and Human3.6M datasets only (second row), and abandoning MoSh parameters (third row). All the different versions have comparable results.

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

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 3

[3] Rıza Alp Güler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, 2019. 3

Figure 3: **Successful reconstructions (1).** We visualize the reconstructions of our approach from different viewpoints.

Data	MoSh	Reconst. Error
All data	Yes	48.6
COCO + H36M	Yes	50.5
COCO + H36M	No	51.4

Table 4: **Ablative on Human3.6M dataset**. We focus on the ResNet backbone and the SMPL head (i.e., we use ground truth bounding boxes) and we ablate different training strategies; using all training data (first row), reducing the training data to COCO and Human3.6M datasets only (second row), and abandoning MoSh parameters (third row). The different versions have comparable results. The numbers are mean 3D joint errors in mm after Procrustes alignment (Protocol 2).

[4] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2

[5] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2013. 2, 3

[6] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2

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

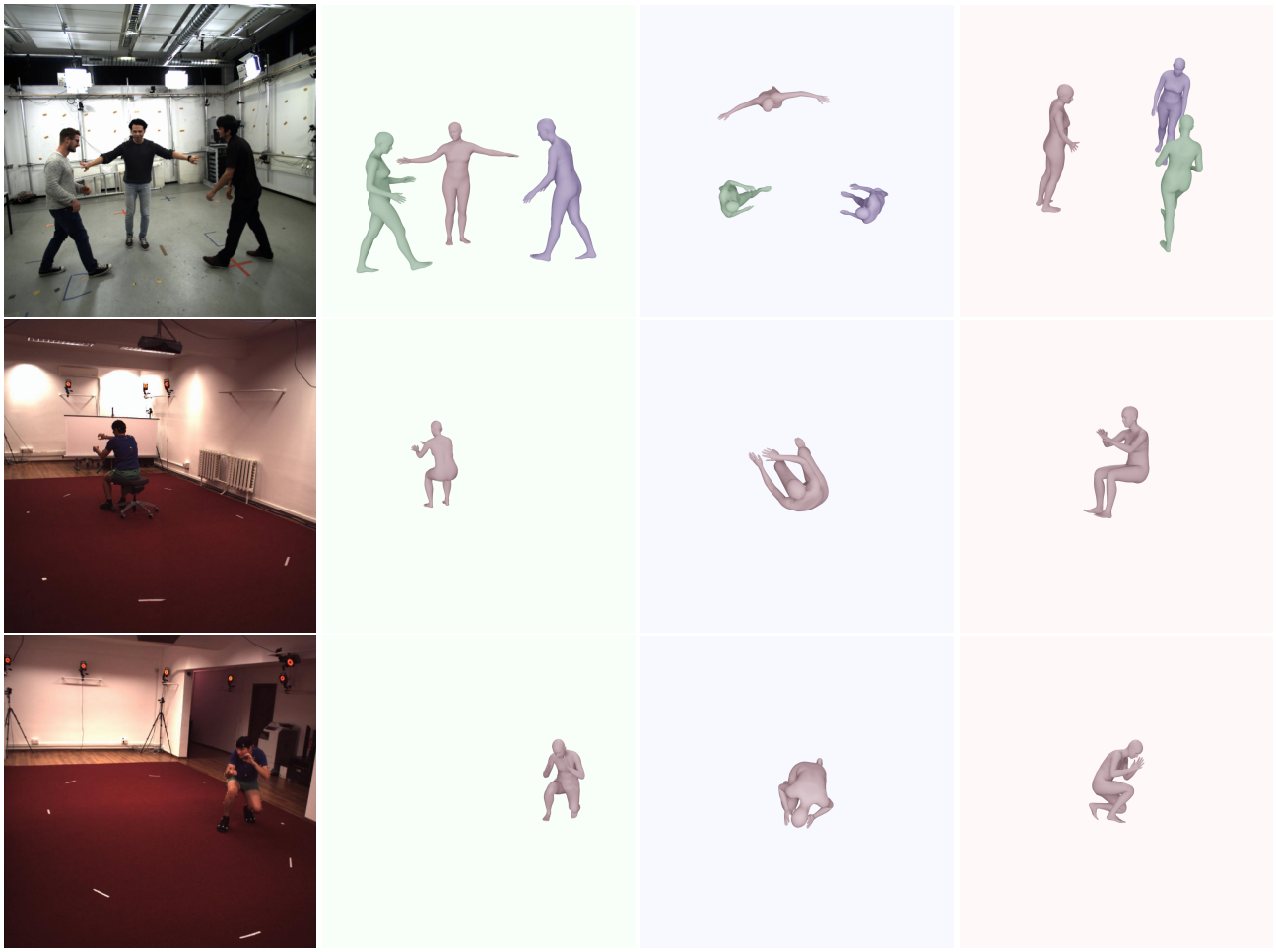


Figure 4: **Successful reconstructions (2).** We visualize the reconstructions of our approach from different viewpoints.

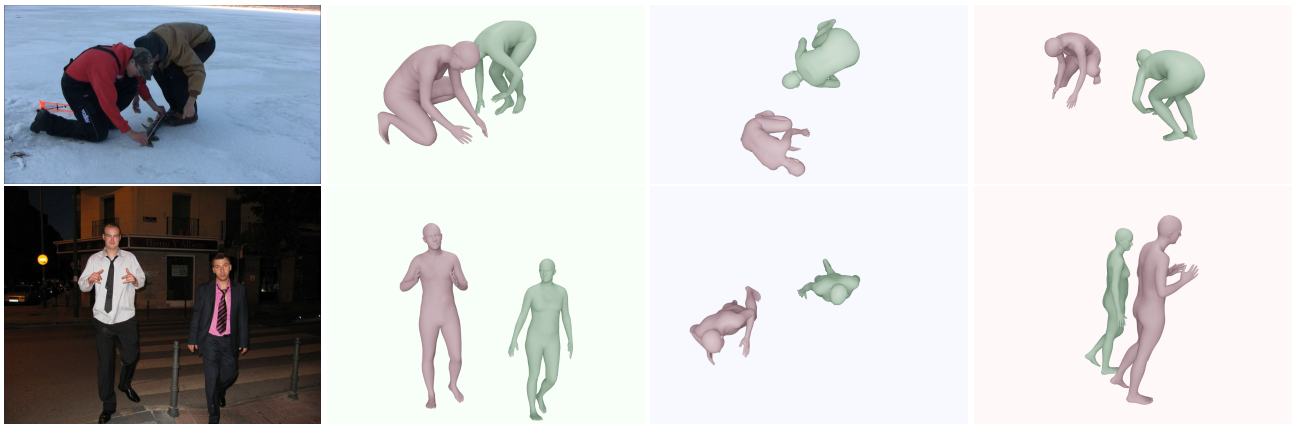


Figure 5: **Failure cases.** We visualize the reconstructions of our approach from different viewpoints. For the first image, the person on the right is slightly shorter than the person on the left, but this is hard to perceive by our model, that estimates roughly the same height for both people and positions the person on the right to be farther away from the camera. For the second image, our model estimates the depth ordering correctly, but clearly overestimates the distance between the two people, which are almost in contact.

2011. 2

[8] 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 *ICCV*, 2015. 3

[9] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe,

Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *PAMI*, 41(1):190–204, 2017. 3

[10] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 3

[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. 3

[12] Abhijit Kundu, Yin Li, and James M Rehg. 3D-RCNN: Instance-level 3D object reconstruction via render-and-compare. In *CVPR*, 2018. 2

[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. 2, 3

[14] 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 preprint arXiv:1908.03265*, 2019. 2

[15] Matthew Loper, Naureen Mahmood, and Michael J Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014. 3

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015. 3

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

[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 *3DV*, 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 *3DV*, 2018. 3

[20] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In *ICCV*, 2019. 3

[21] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018. 3

[22] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 3

[23] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 3

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[25] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net: Localization-classification-regression for human pose. In *CVPR*, 2017. 3

[26] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D pose detection in natural images. *PAMI*, 2019. 3

[27] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 3

[28] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *ICCV*, 2019. 3