Human-Aware Object Placement for Visual Environment Reconstruction **Supplementary Material**

Hongwei Yi[1] Chun-Hao P. Huang[1] Dimitrios Tzionas[1] Muhammed Kocabas[1,2]
Mohamed Hassan[1] Siyu Tang[2] Justus Thies[1] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2] ETH Zürich
{firstname.lastname}@{tuebingen.mpg.de,inf.ethz.ch}

In this supplemental document, we provide additional information about datasets, implementation details, extended sensitivity analysis, failure cases, additional qualitative results and discussion of potential misuse.

1. Dataset

PiGraphs. PiGraphs [7] consists of 60 RGB-D videos of 30 scenes. The dataset is recorded with a *Microsoft Kinect One*, and is designed to capture human and object arrangements in different kinds of interaction. Each video recording is about 2-minute long with 5 fps. It contains labeled 3D bounding boxes of objects in the scene and human poses represented as 3D skeletons. We use this dataset to evaluate the scene reconstruction and compare with [5, 8]. Note that the provided human poses are noisy and not suitable for an evaluation of 3D human shape and pose estimation.

PROX Qualitative. PROX *qualitative* contains 61 RGB-D videos at 30 fps of human motion/interaction in 12 scanned static 3D scenes. The data has been recorded using the *Microsoft Kinect One* and *StructureIO* sensor. To enable 3D scene reconstruction evaluation on this dataset, we segment and label each object with its 3D bounding box. Since there are two scenes (i.e., "BasementSittingBooth" and "N0SittingBooth") containing an inseparable object, we evaluate all methods on the remaining 10 scenes (see Fig. R.1) using the corresponding 51 videos as input.

PROX Quantitative. PROX *quantitative* captures a sequence of human-scene interaction RGB-D frames within a synchronized *Vicon* marker-based motion capturing system. In total, the dataset contains 178 frames and provides groundtruth body meshes, which accounts for human pose and shape (HPS) evaluation. For fair evaluation on HPS, we input all images into HolisticMesh [8] and ours to get a refined scene and use a refined scene to get refined bodies. In addition, we also label this scene for 3D scene reconstruction evaluation, see Fig. R.1.

2. Implementation Details

Loss Terms. The *2D bounding box term* $\mathcal{L}_{\text{bbox}}$ is an ℓ_1 norm between an object's projected 3D bounding box $Proj_i$ and its detected 2D bounding box Det_i, expressed with the top-left corner coordinate x_{min}, y_{min} and $width$ value.

$$\mathcal{L}_{\text{bbox}} = \sum_i \|Proj_i^{\alpha} - Det_i^{\alpha}\|, \quad \alpha \in \{x_{min}, y_{min}, width\}.$$

The *scale term* prevents object scales s deviating far from the initial estimates s^{init} from Total3D [5]:

$$\mathcal{L}_{\text{scale}} = \sum_i \|\frac{s_i}{s_i^{init}} - 1.0\|_2.$$

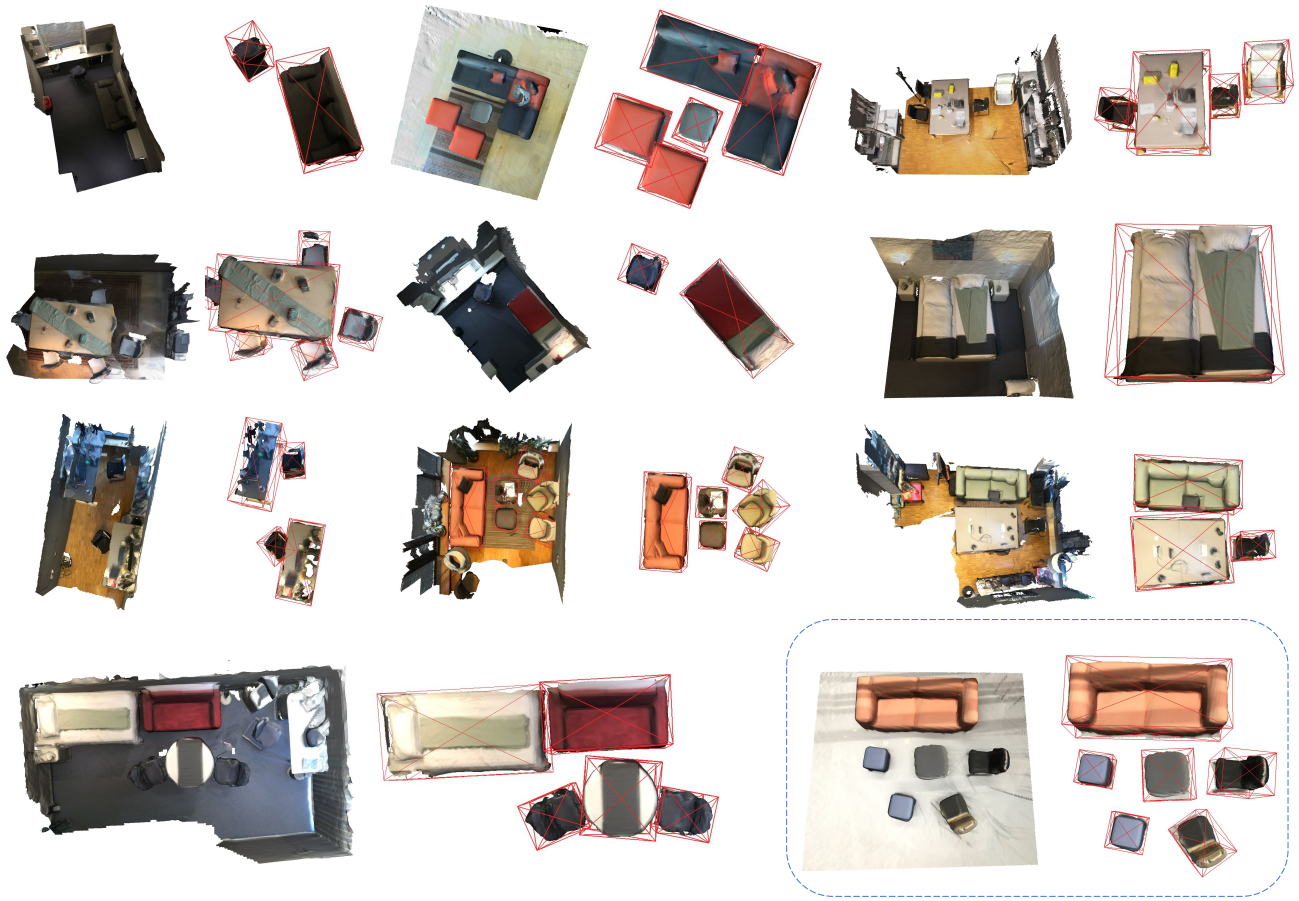
Initial Estimate of 3D Bodies. We use PARE [4] to initialize the body poses and shape (shape β, pose θ, scale s). Since our approach uses the SMPL-X [6] model, we apply [1] to convert the SMPL parameter estimated from PARE. In addition, we use perspective projection with the calibrated camera intrinsic parameters, K provided by the datasets (PiGraph and PROX). To convert the estimations of PARE using a weak perspective camera model, we compute the corresponding translation t^{body} by:

$$\Pi_{K_0}\left(s(R_\theta(J(\beta))\right) = \Pi_K\left((R_\theta(J(\beta)) + t^{body}\right),$$

where K_0 denotes the camera intrinsic parameters of the weak perspective camera model with focal length 5000. Then we extract the resulting 3D joints to initialize E_{body}.

Contact Regions of Objects. We automatically calculate the contact regions of objects based on the normal of the vertices. Specifically, the vertices, whose normals are along y-axis, are the bottom or top part of the objects, while the vertices with along z-axis normal are the back part of the objects. We term that sofas and chairs have two contact regions, i.e., bottom and back parts, while beds and tables only have the top part as the contact region, shown in Fig. R.2.

Optimization. We use the Adam optimizer [3] to optimize the final energy term with a step size of 0.002 and 3000



(A) PROX *qualitative* dataset (B) PROX *quantitative* dataset

Figure R.1. We crop out each object separately and label the corresponding 3D bounding box for 10 scenes in PROX *qualitative* dataset and one scene in PROX *quantitative* dataset.

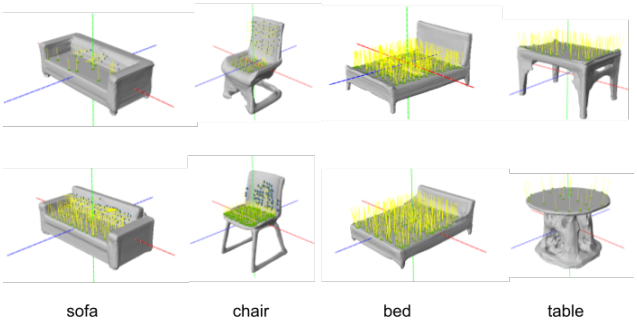


Figure R.2. Contact regions of different objects.

iterations. We set $\lambda_1, \lambda_2, \lambda_3$ as $1000, 0.3, 1000$ respectively, for 2D bounding box term, occlusion-aware term and scale term. The weights of our proposed depth order constraint, collision constraint, and contact constraint are set to $\lambda_4 = 8, \lambda_5 = 1000$, and $\lambda_6 = 1e5$, respectively.

Our method takes around 30 minutes for 3000 iterations to optimize a 3D scene with accumulated HSIs constraints. In comparison, HolisticMesh [8] which jointly optimizes human and a 3D scene for one single image, directly trains the parameters of the network in Total3D [5] to regress the 3D scene, which is time-consuming and costs around 40 minutes. For the human optimization, it runs twice in 5 minutes, i.e., the first pass is a HPS initialization used to refine the scenes, and the second pass is done using the refined scenes. In total, HolisticMesh takes 45 minutes for one single image. Our method takes almost the same time for a scene (around 10 objects) regardless how many frames in the input video. The number of frames in a video only influences the time of calculating the depth map, the SDF volume and the contact information of each body. However, this can be done once and is easily processed in parallel before the optimization. In contrast, HolisticMesh [8] processes a video sequentially, i.e., one frame after another. Therefore, the optimization time increases w.r.t. the number of frames in a video.

3. Sensitivity Analysis.

Our approach uses HSIs observed in a video. A longer video potentially has more HSIs, which results in more constraints for our objective function. In Tab. R.1, we analyze how different video lengths influence scene reconstruction, by reporting the 3D intersection-over-union (IoU) metric. Specifically, we use 10 sequences of the PROX qualitative dataset (one sequence per scene) and randomly sample 10 segments of 10s, 20s, 30s length from each sequence. We observe that longer sequences result in better performance, i.e., higher IoU and lower standard deviation. We observe that the performance of 3D scene reconstruction depends on the *number of HSIs* and not the video length, i.e., a short video with many HSIs results in a better reconstruction than a long video with a few unique HSIs.

	10s	20s	30s	entire videos (51s)
3D IoU mean ↑	0.389	0.395	0.407	0.424
3D IoU std. ↓	0.018	0.015	0.010	-

Table R.1. Ablation study on different length of videos as input. The average length of entire videos is 51s.

We also do a sensitivity study w.r.t. noise in the initialization. In Tab. R.2, we add uniform noise on the initial scale, translation and orientation of objects predicted by Total3D [5], and report the 3D IoU. MOVER is robust to noisy orientation and translation estimates from Total3D [5], but sensitive to the scale variation. This is because we currently regularize the optimization to the initial scale relatively strongly; i.e., we cannot deviate much from a noisy estimate to "correct" it. Relaxing $\mathcal{L}_{\text{scale}}$ easily resolves this.

4. More Evaluation Results on PROX Quantitative Dataset.

We also evaluate 3D scene reconstruction and human-scene interaction on PROX *quantitative*, as shown in Tab. R.3. Our method improves our input baseline [5] significantly and outperforms the previous method [8] with a big margin in both 3D scene reconstruction metrics and human-scene interaction metrics.

5. Failure Cases

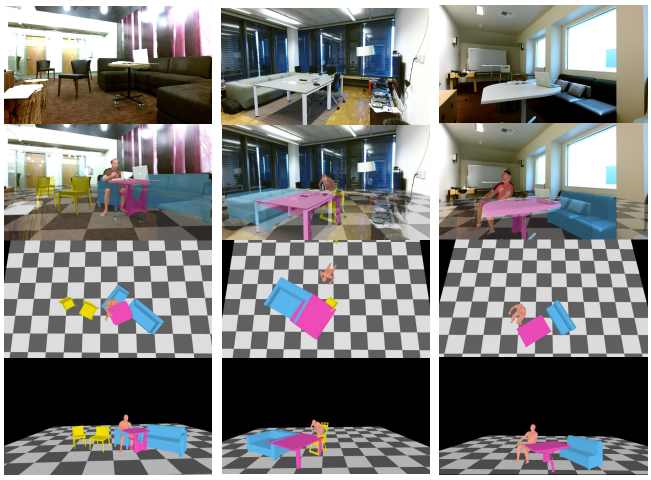
In this section, we discuss and show the failure cases of our method. Besides optimizing the 3D scene layout, we do not change the initial shape estimate of an object. Thus, wrong estimated geometry shape can still violate human's interaction, as shown in (A) in Fig. R.3. A more flexible and adjustable geometry representation, e.g., an implicit representation, would be needed. Human motion reconstruction struggles with severe occlusions in the input, that leads to wrong body poses as well as poor estimations of HSIs, and,

scale noise	± 25%	± 15%	± 0.05%
3D IoU ↑	0.345	0.3805	0.4105
transl.	± 30cm	± 20cm	±10m
3D IoU ↑	0.4175	0.416	0.415
orien.	±45°	±30°	±15°
3D IoU ↑	0.4205	0.418	0.4205

Table R.2. Sensitivity analysis on scene reconstr. with uniform noise on input scale, translation and orientation from Total3D [5] (*Werkraum_03301_01* video). Scene w/o noise has 0.417 3D IoU.

Methods	Scene Recon.			HSI	
	IoU_{3D} ↑	P2S↓	IoU_{2D} ↑	Non-Col ↑	Cont. ↑
HolisticMesh [8]	0.239	0.133	0.533	0.948	0.951
Total3D [5]	0.063	0.409	0.342	0.940	0.436
Ours	**0.390**	**0.095**	**0.862**	**0.972**	**0.934**

Table R.3. Quantitative results for 3D scene understanding (3D object detection) and human-scene interaction on the PROX *quantitative* dataset. P2S, Non-Col and Cont denote *point2surface distance*, Non-Collision and Contactness respectively.



(A) (B) (C)

Figure R.3. Failure cases. (A) The estimated sofa has arms, which does not match the unarmed sofa in the input image. (B) The half bottom body is occluded, that leads to a wrong pose estimation as well as HSI observation. (C) The body is sitting "in the air", where the chair is missing.

thus, influences our 3D scene layout prediction, see (B) in Fig. R.3. While not the scope of our work, the robustness and accuracy of human motion estimation can be improved by incorporating human motion priors or learning-based probabilistic human pose and estimation network. Severe occlusion can also cause missing objects in the scene, like the chair in Fig. R.3(C).

In our pipeline, we currently consider the contact between *detected* objects and bodies. As a potential future extension of our method, one can also leverage the information from

2D learning-based human-object interaction (HOI) detection network [9], by using contacted bodies to discover missing objects; or learn a model that jointly regress human-object interaction and their geometry shape.

6. Additional Qualitative Results

In Fig. R.4 and Fig. R.5, we present additional qualitative results on PROX [2] qualitative and PiGraphs [7] respectively. As can be seen, our method performs well on a variety of different scenes and predicts a physically plausible scene layout. We also refer to the suppl. video for results.

7. Discussion of Potential Misuse

Our approach is not intended for any surveillance application. Our goal is to understand how humans interact and move in scenes from videos (e.g., from TV sitcoms, movies, etc.), to this end both the scene geometry and the human pose and shape need to be reconstructed. Our method could be misused in potential surveillance applications that curtail human rights and civil liberties, but we will restrict the usage of our method in a legal way.

References

[1] https://github.com/vchoutas/smplx/tree/master/transfer_model. 1

[2] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 4

[3] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 1

[4] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11127–11137, 2021. 1

[5] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3DUnderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 55–64, 2020. 1, 2, 3

[6] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1

[7] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning interaction snapshots from observations. *Transactions on Graphics (TOG)*, 35(4):139:1–139:12, 2016. 1, 4

[8] Zhenzhen Weng and Serena Yeung. Holistic 3D human and scene mesh estimation from single view images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 334–343, 2021. 1, 2, 3

[9] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11825–11834, 2021. 4

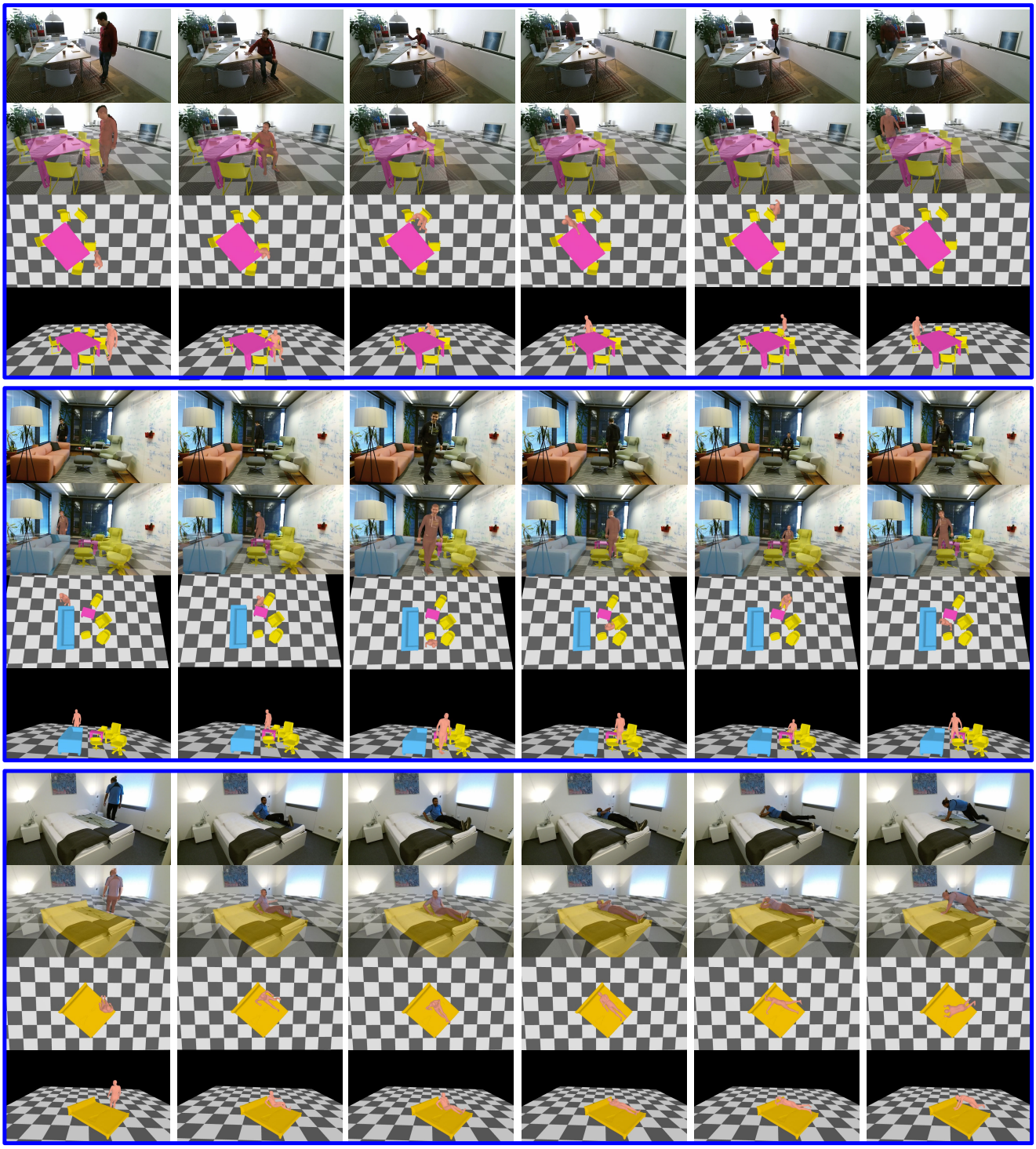


Figure R.4. More qualitative results on PROX qualitative dataset.

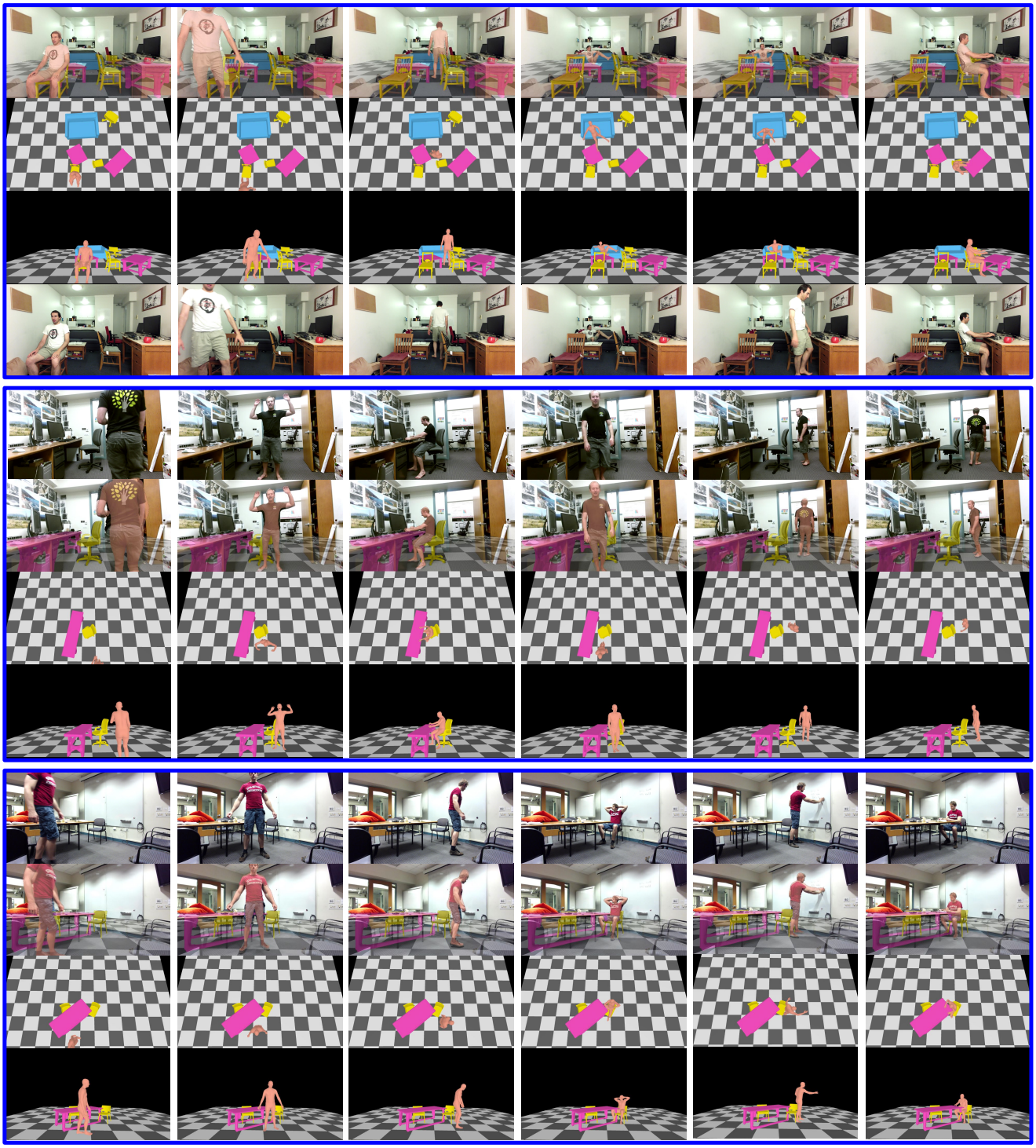


Figure R.5. More qualitative results on PiGraphs dataset.