Supplementary Materials
HSR: Holistic 3D Human-Scene Reconstruction from Monocular Videos

Lixin Xue, Chen Guo, Chengwei Zheng, Fangjinghua Wang, Tianjian Jiang, Hsuan-I Ho, Manuel Kaufmann, Jie Song, and Otmar Hilliges

ETH Zürich, Department of Computer Science
{firstname.lastname}@inf.ethz.ch

In this supplementary document, we provide additional details on data preprocessing (Sec. 1), implementation (Sec. 2), and datasets (Sec. 3). We also present more experiment and ablation results (Sec. 4) to validate our design choices. Lastly, we provide an in-depth discussion of the limitations and possible negative societal impacts of our method (Sec. 5).

In the supplementary video, we illustrate the overall pipeline and showcase video comparisons of our method against baselines.

1 Data Preprocessing

In this section, we detail our data preprocessing procedures. First, we explain the camera localization process in Sec. 1.1 and the human pose estimation in Sec. 1.2. We then describe the alignment of humans and scenes in Sec. 1.3. Finally, Sec. 1.4 and Sec. 1.5 outline the generation of supplementary supervision signals, including monocular cues and masks.

1.1 Scene Preprocessing

Frame Selection. For casually captured videos using a mobile phone, we extract the sharpest frames at equal time intervals to alleviate issues such as motion blur, redundant information from consecutive frames, and excessively long preprocessing time.

Camera Localization. After frame extraction, we perform image undistortion to correct radial distortion, which is the primary type of distortion on mobile phone cameras. We then perform camera localization using COLMAP [23] and HLoc [22]. Specifically, we use the SuperPoint [5] feature combined with LightGlue [16] matching due to their superior accuracy compared to the default SIFT feature and exhaustive matching. To mitigate issues with duplicate structures and repetitive patterns in SfM [2], we restrict feature matching to frames within a temporal window. Figure Fig. 1 illustrates the effects of these engineering efforts. In cases of completely failed localization, we initialize the camera poses with poses from the iPhone's ARKit and perform bundle adjustment for three iterations. In the feature extraction stage, we filter out pixels corresponding to

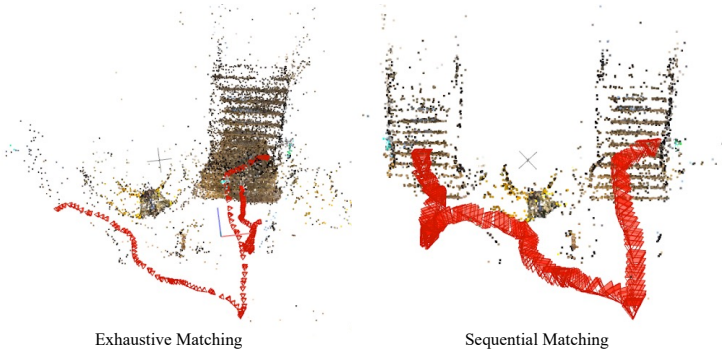

Fig. 1: Camera localization with the presence of duplicate structures. There are two staircases on both sides and the sequential matching can obtain more accurate camera poses with the temporal prior.

dynamic humans using dilated masks from RVM [15]. Lastly, we use BlenderNeuralangelo [14] to extract the bounding regions for the scene normalization. We also obtain sparse point clouds via triangulation from COLMAP, which we use to determine human scales in Sec. 1.3.

1.2 Human Pose Estimation

We utilize SMPL [17] body models as the human prior in human body reconstruction. To estimate the initial SMPL parameters of humans in the videos, we employ the body pose regressor ROMP [24] on each frame separately. The initially estimated per-frame SMPL parameters often exhibit inaccuracies and temporal inconsistencies, particularly in frames where objects occlude humans. Therefore, we further refine the SMPL estimation with a joint loss and a temporal loss following a similar approach to [8]. Specifically, we minimize the 2D distance between the 2D joint predictions $\hat{J}$ obtained from OpenPose [3] and the 2D projection of the 3D SMPL joints $J(\boldsymbol{\theta})$ given SMPL parameters $\boldsymbol{\theta}$:

$$\mathcal{L}_{\text{joint}}(\boldsymbol{\theta}) = \sum_{i=1}^{N_{\text{frame}}} \sum_{j=1}^{N_{\text{joint}}} w_j^i \rho \left(\Pi \left(J(\boldsymbol{\theta}^i)_j \right) - \hat{J}_j^i \right), \tag{1}$$

where Π represents the 3D to 2D projection operators given camera poses and w_i denotes the corresponding OpenPose confidence. To make the optimization robust to outliers, we apply the robust Geman-McClure function $\rho(\cdot)$ on the error term.

In addition, we incorporate a temporal loss that penalizes the difference in the 3D joint locations between consecutive frames:

$$\mathcal{L}_{\text{temp}}(\boldsymbol{\theta}) = \sum_{i=2}^{N_{\text{frame}}} \sum_{j=1}^{N_{\text{joint}}} \left\| J(\boldsymbol{\theta}^i)_j - J(\boldsymbol{\theta}^{i-1})_j \right\|_2^2. \tag{2}$$

This loss penalizes pose jittering and infills plausible body poses in occluded frames.

In this refinement stage, we utilize all the frames in the video for better temporal consistency. We use Adam [11] optimizer to optimize all SMPL parameters for 150 iterations with the aforementioned losses:

$$\mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{joint}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}, \tag{3}$$

where λ_{temp} is a hyperparameter to balance smoothness and fidelity to 2D observations. The effect of such optimization is illustrated in the second and third columns in Fig. 3.

1.3 Human-Scene Alignment

The training of HSR requires the human and the scene to be posed in a globally consistent coordinate system. However, the estimated per-frame SMPL parameters describe the human bodies in the camera coordinate system. Besides, off-the-shelf human pose estimators, such as VIBE [13] and ROMP [24], adopt a weak-perspective camera model based on the assumption that the depth of human bodies is negligible compared to the height and width of human bodies. To align all SMPL estimations in a globally consistent world coordinate system, we need to compute the relative human scale as well as the global 6-DoF root-body trajectories with a perspective camera model.

For the scale ambiguity problem as shown in the left fpart of Fig. 2, we detect the ground plane in the scene and assume the human body will have contact with the ground plane in most of the frames. We transform the SMPL parameters into the world space with the camera poses from COLMAP and compute the minimum scale that one of the world space SMPL mesh vertices will be in contact with the ground plane for each frame. We take the median of the estimated scales from all the frames and use it as the single scale for all the frames. The right part of Fig. 2 shows the results after alignment. This helps us reduce the temporal jittering in scales and makes our method robust to certain human motions such as jumping. In case of failed ground detection due to few feature detections in the textureless ground, our pipeline allows for a user-specified scale that visually aligns humans with the scene and can optimize the scale in the joint optimization in the later stage.

Once a reasonable scale estimation is obtained, we transform the SMPL parameters from the camera coordinate frame to the world coordinate frame with the following transformation [19]:

$$R_w = R_c^w R_c, \tag{4}$$

$$t_w = R_c^w(p + t_c) + t_c^w - p \tag{5}$$

where p is the pelvis location for the estimated shape parameters, R_c is the global orientation in the camera coordinate frame, t_c the translation in the camera

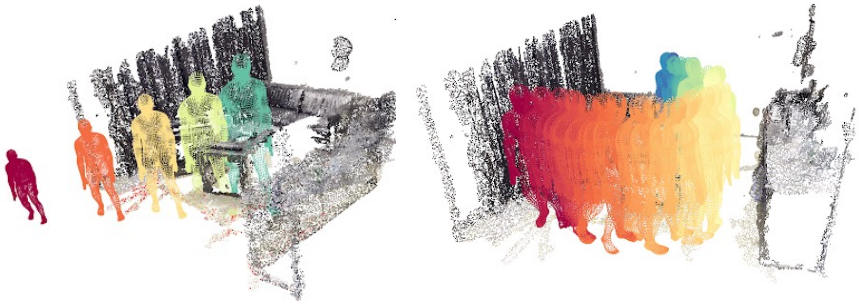

Fig. 2: Human-Scene Alignment. We use the contact prior to resolve the scale ambiguity (left) and obtain a globally coherent human pose estimation in the world coordinate system (right).

coordinate frame, (R_w, t_w) the SMPL parameters in the world coordinate frame, and (R_c^w, t_c^w) the transformation from the camera coordinate frame to the world coordinate frame.

1.4 Monocular Geometric Cues

We use pretrained models from Omnidata [7] to predict monocular depth and normal for each frame. The model is trained on images of size 384×384, so we take patches of this size from the original image to estimate the corresponding cues and align different patches together based on the overlapping regions. The estimated depth is not metric and subjects to a scale and shift transformation. Therefore, we apply a least square estimator to solve the scale and shift needed to align patches of depth maps. For surface normals, we estimate the least square solution of the rotation matrices to align patches of normal maps. This patch-based monocular cue estimation allow us to obtain monocular cues for arbitrary high-resolution input without the need for resizing or cropping.

1.5 Human Masks

We use SAM [12] for additional supervision on the foreground-background separation. With proper prompting, SAM masks are accurate and robust to occlusion. We initialize the prompt based on the mask from a human video matting model RVM [15] and the SMPL body mask. The SMPL body mask is clean but does not take into account occlusion. The RVM mask considers occlusion but may have random noise in some regions of the image. Therefore, we use the intersection of these two masks as the initial mask prompt for the SAM module. Additionnally, we provide the bounding box of the intersection mask and 2D keypoints from OpenPose as prompts. The initial SAM mask can be noisy at the boundaries of the human mask. To address this, we use the estimated mask and sampled random points from the mask as the new prompt for the SAM module. We iterate

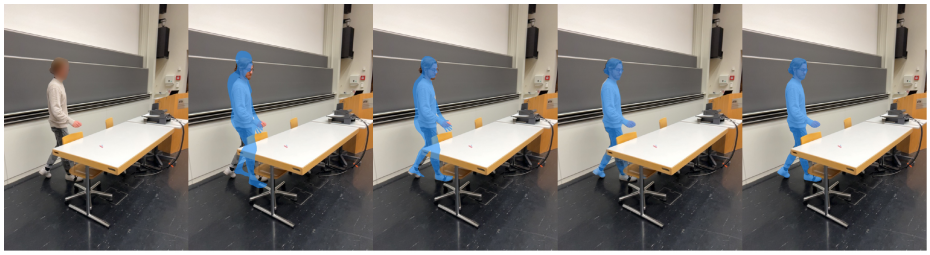

Fig. 3: Human Masks. We show how we gradually obtain accurate human masks based on the noisy masks from other modules.

this process for 30 iterations to obtain a clean and occlusion-aware human mask to assist in human-scene decoupling. The initial SMPL mask and RVM mask are shown in the third to fifth columns of Fig. 3 for visual comparison.

2 Implementation Details

2.1 Network Architecture

Human. The canonical human shape network f^H_{sdf} (Eq. 1 in the main manuscript) comprises 8 blocks. Each block consists of a fully connected layer, a weight normalization layer [21], and a softplus activation layer [6]. The fully connected layer contains 256 neurons. The pose condition, denoted as $\boldsymbol{\theta}$, is obtained by concatenating all axis angles represented in radians. To better model high-frequency details, we apply positional encoding with 6 frequency components to the input points, following the approach presented in [18]. The canonical human texture network, denoted as f^H_{rgb} (Eq. 4 in the main manuscript), consists of four blocks. These blocks share the same architecture as the human shape network, except using the Sigmoid activation function for the last layer and ReLU activation functions for the remaining layers. To expedite the convergence of the model, we employ a pretraining strategy for the shape network. Specifically, we initialize the shape network using a SMPL mesh in the canonical pose.

Scene. The scene shape network, denoted as f^S_{sdf}, and the scene texture network, denoted as f^S_{rgb} (Eq. 2 and Eq. 5 in the main manuscript), exhibit similar network architectures to the canonical human network design, except that we have only 2 blocks for the scene texture network. The scene texture network incorporates the view direction $\mathbf{v}$ and a per-frame learnable frame code as additional input conditions, replacing the human pose. The learnable frame code is used to take into account the dynamic shadows caused by the moving humans in the scene. We deploy positional encoding with 4 frequency components to the view direction to help model the view-dependent photometric effects.

2.2 Training Details

Our networks are trained using the Adam optimizer [11], with an initial learning rate of $5e^{-4}$. This learning rate is exponentially decayed to $5e^{-5}$ when training ends. The other Adam hyper-parameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The complete model is trained for 250k steps, requiring approximately 24 hours of training time on an NVIDIA RTX 4090 GPU. We use a weighted pixel sampling strategy based on the 2D human bounding box derived from the predicted human masks. We put 80% of pixel samples uniformly inside the 2D human bounding box to learn the human models and separate the foreground and background. The remaining 20% of samples are uniformly distributed over the whole image for the background reconstruction.

2.3 Losses

Here we provide more details regarding the depth loss and the mask loss.

Depth Loss. Following MonoSDF [26], we enforce the consistency between our rendered depth map $D(\mathbf{r})$ and the depth map $\hat{D}(\mathbf{r})$ predicted by the pre-trained Omnidata model [7] using a scale-invariant loss term:

$$\mathcal{L}^k_{\text{depth}} = \sum_{\mathbf{r}\in\mathcal{R}^k} \|(w\hat{D}(\mathbf{r}) + q) - D(\mathbf{r})\|^2, \tag{6}$$

where w and q are the scale and shift factor used to align $D(\mathbf{r})$ and $\hat{D}(\mathbf{r})$, and R^k are pixels in batch k. We compute w and q with a least-square criterion [20]:

$$(w, q) = \arg\min_{w,q} \sum_{\mathbf{r}\in\mathcal{R}} (w\hat{D}(\mathbf{r}) + q - D(\mathbf{r}))^2. \tag{7}$$

w and q can be determined analytically using a closed-form solution:

$$\begin{bmatrix} w^* \\ q^* \end{bmatrix} = \left(\sum_{\mathbf{r}} \mathbf{d_r}\mathbf{d_r}^T\right)^{-1} \left(\sum_{\mathbf{r}} \mathbf{d_r} D(\mathbf{r})\right) \tag{8}$$

where $\mathbf{d_r} = (\hat{D}(\mathbf{r}), 1)^T$. In line with the approach presented in [26], we estimate the parameters w and q at each iteration by sampling a batch of rays randomly within a single image. This strategy is adopted because the monocular depth predictor cannot consistently provide accurate scales and shifts across frames due to the varying scene geometry. It is important to note that the monocular cues pertaining to the human subjects lack sufficient accuracy, and therefore, we refrain from using the estimated depth or normal information of the human subjects for supervision.

Mask Loss. The self-supervised scene decomposition loss [9] aims to guide the optimization towards a clean and robust decoupling of the human and the scene. However, this loss alone cannot perfectly decouple the dynamic human and the static environment mainly due to two factors. The first factor is the dynamic

components like shadows which break the color consistency of the static scene across different frames. The other issue is the inaccuracy in human pose estimation. With wrong human poses, the network has to learn the background color in the human model and the human color in the background model. Therefore, we propose to add a foreground and background mask loss to help guide the separation and the refinement of human pose error. Specifically, we apply a L_1 loss on the occlusion-aware foreground accumulated weights and background accumulated weights.

2.4 Volume Integration

There exist multiple ways for volume integration of two neural fields. For example, one could take the minimum of SDF values in each field and use it as the joint SDF value for the location. An alternative way is to obtain the density value from SDF values in each neural field and use the linear property to add two density values together as the joint density value for the input location. We empirically find that these two alternative volume integration schemes deliver inferior results. Besides, these two schemes require the evaluation of points in two neural fields, thus more computationally expensive compared to the scheme proposed in the main paper. As a result, we use separate sets of points for each neural field.

3 Datasets

3.1 SHSD Dataset

We introduce a semi-synthetic dataset called the Synthetic Human Scene Dataset (SHSD) specifically designed for the quantitative evaluation of human-scene reconstruction methods. SHSD comprises six sequences featuring different human subjects moving around in the scene. The sequences present realistic and challenging scenarios with varying degrees of occlusion. Each sequence consists of 75 to 150 frames capturing the same human subject in different poses. These frames are rendered in artist-authored indoor stages sourced from the ReplicaCAD [25] and CIRCLE [1] datasets. The camera trajectories follow an arc around the human subjects, ensuring comprehensive coverage. To make the depth scale consistent between ground truth and our preprocessed data, we align the cameras estimated from COLMAP with the ground truth cameras via a single similarity transformation.

For better visualization, Fig. 4 and Fig. 5 provide bird's eye views of the stages extracted from the ReplicaCAD and CIRCLE datasets. Furthermore, Fig. 6 showcases four example frames from two sequences, each accompanied by corresponding ground truth information such as masks, depth maps, and surface normals.

In addition, we also provide the ground truth 3D human scans for evaluation of the geometry reconstruction quality. The human scans utilized in the proposed SHSD are captured using a multi-view photogrammetry system [4]. This

Fig. 4: BEV of one synthetic stage from ReplicaCAD [25] dataset

Fig. 5: BEV of the synthetic stage from CIRCLE [1] dataset.

system comprises 53 RGB and 53 IR cameras, enabling comprehensive coverage and detailed capture. Each scan consists of a high-resolution 40K-face mesh representing the geometry of the human subject, accompanied by a 4K-resolution texture map that provides appearance information.

It is important to note that the collection of this data, as well as its usage involving human subjects, has undergone a rigorous ethical review and approval process by an Institutional Review Board. This ensures that all necessary ethical considerations and guidelines are adhered to.

We utilize Blender to generate realistic renderings and obtain various types of ground truth information, such as full image depth, surface normals, and human masks. For sequences created in the ReplicaCAD stages, we use lighting-baked texture. For sequences created in the CIRCLE dataset, we use the default lighting provided by the authors and render shaded colors.

3.2 Real Dataset

In our experimentation, we employ an iPhone 13 Pro Max to capture six real sequences featuring individuals performing various movements within different scenes. In certain sequences, we also feature occlusion and close contact with the objects in the scene. Two of these sequences are recorded using the native camera application on the iPhone, simulating real-world scenarios where auto-focus and auto-white-balancing are enabled. This setup introduces challenges as the focal length and white balance may vary throughout the sequences, while our model assumes fixed focal length and color consistency. The remaining eight sequences are recorded using Record3D, which maintains a fixed focal length and white balance for color consistency. Besides, we can easily export ARKit camera poses from Record3D and use it as an initialization in case of failed reconstruction due to duplicate structures.

4 Additional Experiment Results

4.1 Evaluation Deatils

For fair comparison over human reconstructions, we use HSR-processed data as the input for SelfRecon [10] and Vid2Avatar [9] instead of using their own

preprocessing pipelines. As a result, the reconstruction difference from different methods solely comes from the design choices of different algorithms. For quantitative evaluation on novel view synthesis, we leave out one frame from every 10 frames. For foreground comparison, we set non-human pixels to white according to ground truth masks and then report metrics on it following previous works [9, 10]. As for the metrics on geometry reconstruction, we perform ICP between reconstructed meshes and ground truth meshes. This is because raw reconstructed meshes are not well-aligned with the ground truth meshes.

To validate our design choices, we perform a quantitative evaluation on a subset of synthetic sequences. The result is shown in Tab. 1. We will discuss individual columns in the following subsections.

	PSNR ↑	SSIM ↑	LPIPS ↓	IoU ↑	Normal ↑	Depth↓
Full	**27.56**	**0.9379**	**0.0608**	**0.9637**	**0.9678**	**0.1003**
- Shape-aware Sampling	27.25	0.9349	0.0629	0.9619	0.9653	0.1097
- Global Coord.	14.13	0.7347	0.3736	0.9467	0.7946	0.9718
- Depth Loss	27.11	0.9309	0.0745	0.9625	0.9614	0.1340
- Normal Loss	25.66	0.9361	0.0655	0.9623	0.8839	0.1748

Table 1: Quantitative ablation results on SHSD. We demonstrate the importance of the proposed components through metrics on novel view synthesis and geometry.

4.2 Occlusion

Our method handles the human-scene occlusion naturally via the SDF-based compositional volume rendering and the reconstructed surfaces. To further clarify this process, we include an illustration in the second row of Fig. 7. Given the reconstructed foreground objects (occluders) and humans, our inverse CDF sampling encourages a dense ray sample distribution close to the respective surfaces (Fig. 7 (e)). This approach ensures a clear separation, leading to a correct volume integration with accurate color rendering.

Using a globally consistent coordinate system rather than a human-centric coordinate system is necessary for the reconstruction of the scene, thus correctly modeling occlusion. We provide evidence in Fig. 7 (c) that without a globally consistent coordinate system, the 3D human cannot be fully recovered even using geometric cues. Cross-view photometric consistency provided by a global coordinate system is essential for 3D scene reconstruction.

Similarly, without the proposed shape-aware sampling, human samples might scatter across the scene due to the limitation of inverse CDF sampling. As a result, we do not have enough samples around the actual human surface, leading to truncated human bodies as shown in Fig. 7 (d).

4.3 Mask Loss

We ablate the proposed mask loss and qualitatively demonstrate its effectiveness in Fig. 8. Initial SMPL estimations are often inaccurate especially when some

part of the human body is occluded, as shown in the second column in Fig. 8. In contrast, 2D human segmentation is more robust and accurate. Therefore, we employ such mask supervision to help refine large pose errors that other 2D rendering losses have trouble with. As a result, we refine the human pose and learn a better human-scene decomposition as shown in the last two columns in Fig. 8.

4.4 Contacts with other scene elements.

Our method can handle interactions with any element within the scene, including chairs and tables, through our general formulation of implicit scene modeling and the interpenetration loss. We provide a detailed examination of a scenario where a human subject is sitting on a chair in Fig. 9. Such contacts between the upper leg and the chair are invisible in the input image, but our interpenetration loss and scene modelling lead to a plausible and penetration-free reconstruction.

4.5 Human Reconstruction Quality

We show more comparison on human reconstruction quality in Fig. 10. Compared to Vid2Avatar, our method is more robust to initial pose error as shown in the third, fourth, and sixth rows in Fig. 10. Compared to SelfRecon which requires human masks as input and cannot recover from wrong masks, our method separates foreground and background more clearly and has much better rendering quality and geometry reconstruction. Moreover, both SelfRecon and Vid2Avatar assume there is no occluder in front of the human body. Therefore, they cannot handle occlusion at all. As shown in the fifth row in Fig. 10, Vid2Avatar learns the wrong texture on the human body and SelfRecon completely distorts the lower part of the human body.

4.6 Holistic Reconstruction Quality

Fig. 11 shows the geometry reconstruction of different methods on four real sequences. Our result demonstrates superior quality for both foreground humans and background scenes compared to PPR and Total-Recon. It is worth noting that TotalRecon requires LIDAR depth as input, while still producing worse reconstruction compared to ours. Moreover, the reconstruction from our method preserves the correct spatial relationship between humans and scenes due to our explicit enforcement of physical constraints. In contrast, PPR only fits a ground plane and performs physical simulation on the ground plane, and TotalRecon does not take human-scene interaction into account. As a result, complex interactions like touching the wall (2nd row) and lying on the bed (4th row) are not correctly modeled by the baseline methods.

5 Limitations and Negative Societal Impact

Limitations. Our framework is limited to reconstructing a single person from the video. To reconstruct multiple persons in the scene, it is straightforward

to apply multiple human fields in our framework. However, it is challenging to accurately estimate the poses of multiple humans, especially under close interaction. In the current framework, we also assume that the scene is static. To make the method applicable for general dynamic scenes, such as videos involving human-object interactions, is another interesting direction. This requires accurate estimation of hand poses and object poses, which is hard as we only have low-resolution observations of the human hands and the corresponding objects. We leave these two extensions as future works.

Negative Societal Impact. HSR facilitates the conversion of humans and scenes into digital forms through a single RGB video, offering vast possibilities for the film industry, augmented and virtual reality, and telepresence applications. Our technique produces a digital human avatar that can be animated to adopt previously unseen poses. However, there exists a potential for misuse, such as the creation of deep-fakes. Addressing these concerns is paramount before integrating digital human avatars into products. Our intention is to foster applications of this technology that benefit society. Although it is impossible to completely eliminate the risk of malicious use, we believe that conducting research with maximum transparency—by openly discussing technical details, and sharing code and data—is the best approach. This openness is not only ethical but also instrumental in developing countermeasures against misuse, thus mitigating the risks associated with dubious applications.

References

1. Araújo, J.P., Li, J., Vetrivel, K., Agarwal, R., Wu, J., Gopinath, D., Clegg, A.W., Liu, K.: Circle: Capture in rich contextual environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21211–21221 (2023) 7, 8
2. Cai, R., Tung, J., Wang, Q., Averbuch-Elor, H., Hariharan, B., Snavely, N.: Doppelgangers: Learning to disambiguate images of similar structures. In: ICCV (2023) 1
3. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 2
4. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Trans. Graph. **34**(4) (jul 2015). https://doi.org/10.1145/2766945, https://doi.org/10.1145/2766945 7
5. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE (Jun 2018). https://doi.org/10.1109/cvprw.2018.00060, http://dx.doi.org/10.1109/CVPRW.2018.00060 1
6. Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., Garcia, R.: Incorporating second-order functional knowledge for better option pricing. In: Leen, T., Dietterich, T., Tresp, V. (eds.) Advances in Neural Information Processing Systems.

vol. 13. MIT Press (2000), https://proceedings.neurips.cc/paper/2000/file/44968aece94f667e4095002d140b5896-Paper.pdf 5
7. Eftekhar, A., Sax, A., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10786–10796 (2021) 4, 6
8. Guo, C., Chen, X., Song, J., Hilliges, O.: Human performance capture from monocular video in the wild. In: 2021 International Conference on 3D Vision (3DV). pp. 889–898. IEEE (2021) 2
9. Guo, C., Jiang, T., Chen, X., Song, J., Hilliges, O.: Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12858–12868 (2023) 6, 8, 9
10. Jiang, B., Hong, Y., Bao, H., Zhang, J.: Selfrecon: Self reconstruction your digital avatar from monocular video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 8, 9
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2015) 3, 6
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023) 4
13. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5253–5263 (2020) 3
14. Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2
15. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance (2021) 2, 4
16. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local Feature Matching at Light Speed. In: ICCV (2023) 1
17. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015) 2
18. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020) 5
19. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019) 3
20. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022) 6
21. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016), https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf 5

22. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: CVPR (2019) 1
23. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1
24. Sun, Y., Bao, Q., Liu, W., Fu, Y., Michael J., B., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: ICCV (2021) 2, 3
25. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., Batra, D.: Habitat 2.0: Training home assistants to rearrange their habitat. In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 7, 8
26. Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in Neural Information Processing Systems (NeurIPS) (2022) 6

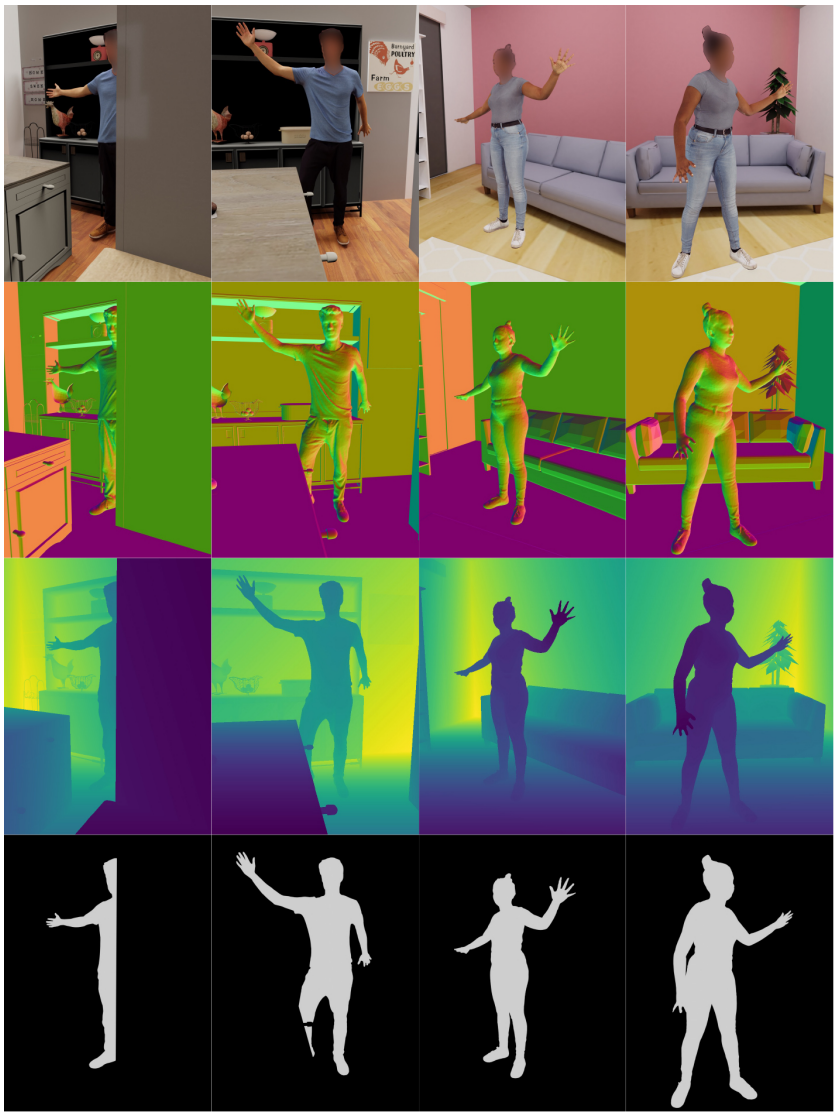

Fig. 6: Sample ground truth images from SHSD. From top to bottom: RGB, normals, depths, and masks.

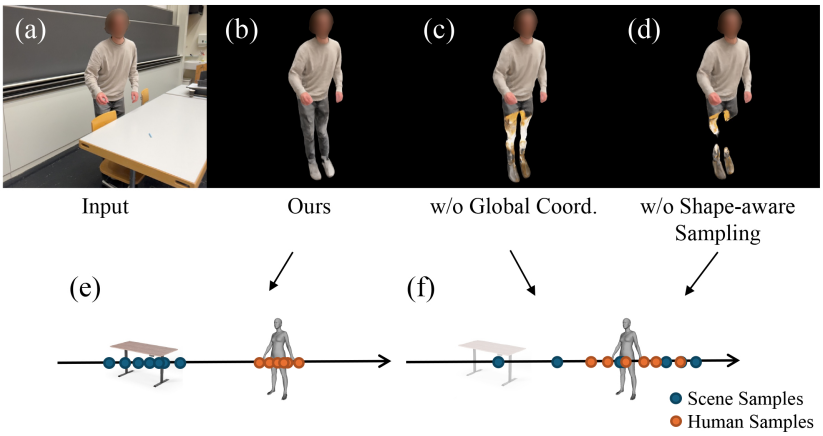

Fig. 7: Illustration for occlusion handling and ablation results.

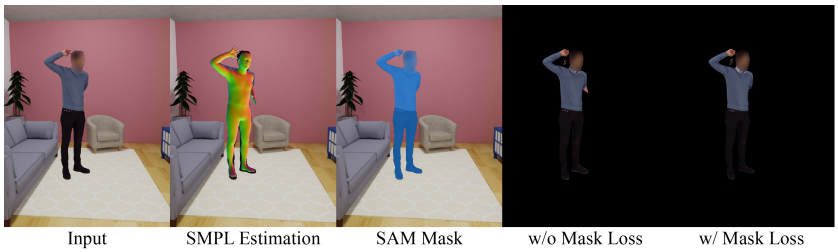

Fig. 8: Effectiveness of mask loss. Robust and accurate 2D masks can help correct large initial pose estimation errors.

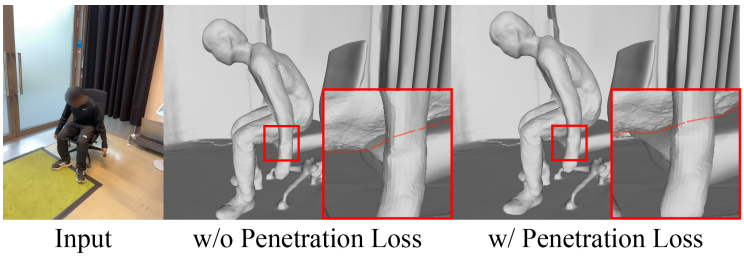

Fig. 9: Human-Chair contacts. The red dashed lines indicate the human mesh outlines.

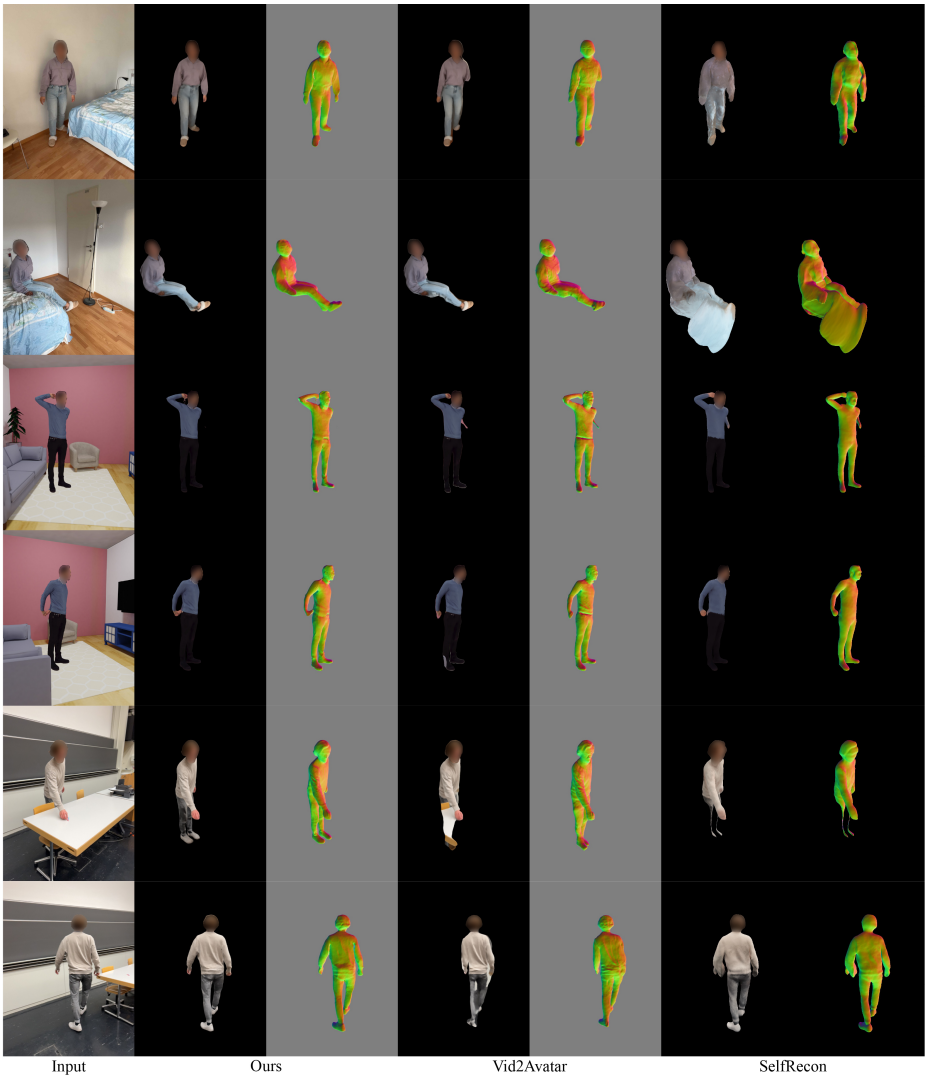

Fig. 10: Qualitative results on human reconstruction and rendering. Our method shows better human-scene decomposition and robustness to pose error and strong occlusion compared to baselines.

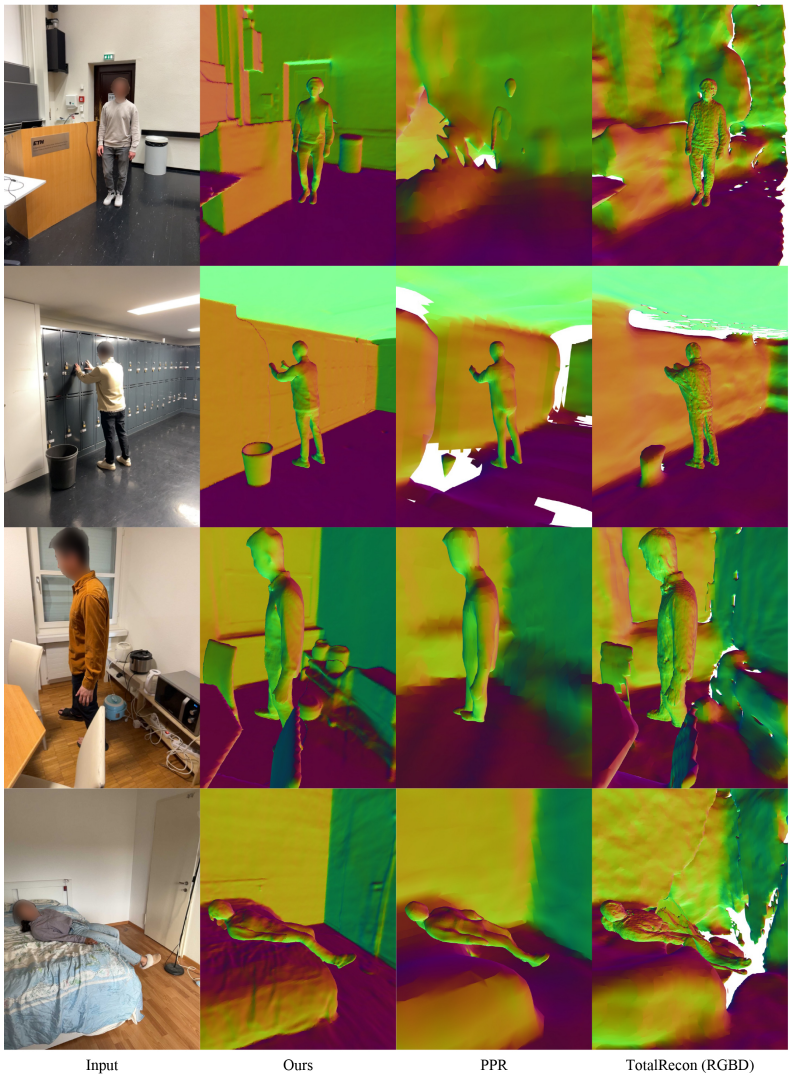

Fig. 11: Qualitative results on holistic reconstruction. Our method has much better foreground and background reconstruction compared to baselines. As a result, we correctly model the human-scene contact while baselines have severe interpenetration.