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

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

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

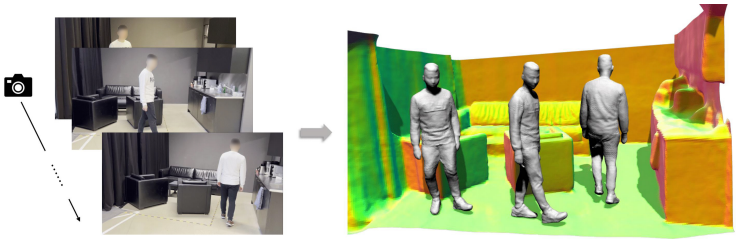

Fig. 1: HSR jointly reconstructs dynamic humans and static scenes in a shared global coordinate system from monocular RGB videos. This approach enables HSR to account for physical interactions between humans and scenes, effectively addressing issues of interpenetration and occlusion.

Abstract. An overarching goal for computer-aided perception systems is the holistic understanding of the human-centric 3D world, including faithful reconstructions of humans, scenes, and their global spatial relationships. While recent progress in monocular 3D reconstruction has been made for footage of either humans or scenes alone, the joint reconstruction of both humans and scenes, along with their global spatial information, remains an unsolved challenge. To address this, we introduce a novel and unified framework that simultaneously achieves temporally and spatially coherent 3D reconstruction of static scenes with dynamic humans from monocular RGB videos. Specifically, we parameterize temporally consistent canonical human models and static scene representations using two neural fields in a shared 3D space. Additionally, we develop a global optimization framework that considers physical constraints imposed by potential human-scene interpenetration and occlusion. Compared to separate reconstructions, our framework enables detailed and holistic geometry reconstructions of both humans and scenes. Furthermore, we introduce a synthetic dataset for quantitative evaluations. Extensive experiments and ablation studies on both real-world and synthetic videos demonstrate the efficacy of our framework in monocular human-scene reconstruction. Code and data are publicly available on our project page.

* Corresponding author, now at HKUST(GZ) & HKUST.

Keywords: Neural implicit representations · 3D human body shape modeling · Scene reconsturction

1 Introduction

Digitally recreating real-life scenarios, particularly those involving human-centric activities, is crucial for empowering machines to perceive and interact with the world around them. Consider a delivery robot navigating to hand over a package to a person as a motivational example. For successful interaction, the robot's perception system must perceive the 3D layout of its surroundings and accurately interpret the state and dynamics of the humans within that space. This necessitates a comprehensive approach to reconstructing both humans and scenes, along with understanding their spatial relationships. Moreover, this reconstruction capability must be robust and capable of adapting to new individuals, varying clothing styles, and diverse environments without specific templates. Therefore, our goal is to achieve holistic 3D reconstructions of static scenes and dynamic human occupants from monocular RGB videos captured by consumer devices.

Most previous works treat static scene reconstruction and dynamic human reconstruction separately. For example, with advances in neural volume rendering [44, 60, 70, 71], researchers have explored reconstructing scenes with neural implicit functions [40, 48, 59, 74], but all of these works ignore dynamic components such as humans. Simultaneously, there has been remarkable progress in deformable object reconstruction from images, especially articulated humans [6, 21, 24, 37, 50, 65]. Although successful in their respective tasks, the challenge remains in leveraging these methods to simultaneously reconstruct both scenes and humans as motivated above. As we have found experimentally, naively combining these two lines of methods produces artifacts such as truncated human bodies and human-scene interpenetration This is because it lacks the spatial relationship between humans and scenes, where humans are unaware of scenes and vice versa. In addition, most human reconstruction methods assume a complete observation of the human, ignoring the common human-scene occlusion.

To address these issues, we propose HSR based on the following insight: modeling static scenes and dynamic humans should be reconciled into a single, unified framework that treats the problem holistically. More concretely, we utilize two neural fields - one for modeling the dynamic human in canonical space [21] and another for the static scene [74]. We formulate a global, joint optimization over all of the learnable parameters of the dynamic human and static scene. This includes physical constraints imposed by the potential human-scene occlusion and interpenetration. In addition, a 3D body model-guided sampling strategy for surface-based neural volume rendering then facilitates the separation of the dynamic human and static environment. The 3D body model is utilized to guide the sampling process, which helps to effectively model sharp boundaries between the dynamic human and static environment. This proves beneficial even when faced with significant human-scene occlusions. We also use estimated normal and

depth maps to improve reconstruction quality. This approach enables plausible and detailed holistic geometry reconstructions of the entire space

Our experiments demonstrate that our method achieves accurate human-scene decomposition and detailed 3D reconstruction. Furthermore, comparisons with existing methods indicate superior performance of our method compared to prior works. To facilitate quantitative comparison, we contribute a novel semi-synthetic test set with accurate 3D geometry of complete scenes. Finally, extensive ablation experiments validate our design choices.

In summary, our contributions are:

- a novel unified framework to holistically reconstruct 3D scenes with dynamic people from monocular videos, achieving robust and detailed 3D reconstructions with a clean separation even under challenging human-scene occlusions and interpenetrations,
- a novel semi-synthetic dataset with rich 3D annotations, allowing for comparing monocular dynamic scene reconstruction methods,
- extensive ablation studies and comparisons demonstrating the effectiveness of our proposed components and the entire system.

2 Related Work

Dynamic Human Reconstruction. Explicit representations have been successfully used for monocular human performance capture but required personalized, manually rigged templates, such as human scans [24, 25, 66]. Some methods [1, 8, 20, 46] have addressed this, but explicit mesh representations are inherently limited by a fixed resolution. Methods that directly regress 3D surfaces have shown impressive results [2, 15, 27–29, 32, 49, 50, 65, 79]. However, they often struggle with building a consistent representation over time and require 3D data for supervision. More recently, implicit neural fields combined with neural rendering have emerged as a way to fit articulated human models to videos [21, 33–35, 39, 47, 54, 55, 63]. Vid2Avatar [21] and SelfRecon [33] improve geometry reconstruction quality by adopting an SDF-based representation and enable reconstructing human avatars from monocular videos. OccNeRF [64] incorporates attention and graph convolution to hallucinate occluded regions and achieve plausible novel view rendering. Unlike these methods that focus primarily on the human subject, our work aims to achieve a consistent 3D reconstruction of both humans and scenes.

Static Scene Reconstruction. Multi-view stereo (MVS) is a common technique for reconstructing the dense geometry of static scenes [17, 19, 52, 58, 68, 69]. Many learning-based MVS methods leverage CNNs to overcome the limitation of hand-crafted modeling attempts of traditional methods [9, 19, 42, 57, 58, 68, 69]. Recently, with the rise of NeRF [44], researchers have explored static scene reconstructions with neural implicit functions. IDR [71] reconstructs surfaces as the zero-level set of an MLP. While achieving good reconstruction, IDR requires accurate object masks during training. To avoid using masks, VolSDF [70] and NeuS [60] modify NeRF and use signed distance function (SDF) to represent the

density function in neural volume rendering. Since some works [62,75] show that there exists an ambiguity between appearance and geometry in NeRF, many variants based on VolSDF or NeuS adopt better geometry supervision and improve reconstruction accuracy [13, 16, 59]. MonoSDF [74] uses monocular depth and normal estimation to provide geometric constraints across images. However, all these methods consider only static scenes without moving subjects, limiting their applicability to life-like scenes.

Human-Scene Interaction. Human-scene interaction in 2D and 3D has been widely studied in the literature [5, 7, 22, 26, 30, 31, 38, 45, 51, 72, 73]. A line of research estimates human body poses in scenes from RGB images [26,30] or with the aid of wearable sensors [12,23,36,43,77]. Most methods require a prescanned scene and do not estimate detailed human surface geometry. Recently, TotalRecon [53] focuses on reconstructing articulated objects and scenes from RGBD input. PPR [67] reconstructs articulated objects and scenes from monocular videos, employing physical simulation to determine relative scales and optimize poses. However, they do not model humans explicitly, leading to insufficient reconstruction quality of human subjects. In summary, prior work either assumes a ready-made 3D scene scan, does not reconstruct the detailed human surface geometry, or requires additional sensor inputs. These limitations prevent such techniques from being deployed to applications where prior knowledge of scenes and human surface geometry cannot be afforded.

3 Method

We first describe how we model the geometry and appearance for the human and the scene in Sec. 3.1. Next, we discuss how we obtain a final pixel value via compositional volume rendering in Sec. 3.2. Finally, Sec. 3.3 shows our global optimization procedure. For an overview of our method, please refer to Fig. 2. In the following, we assume that two sets of 3D points have been sampled along a ray: one set for the human part $\{\mathbf{x}_d\}_{i=1}^N$, and the other for the scene $\{\mathbf{x}_s\}_{i=1}^N$. More details on sampling are explained in Sec. 3.2.

3.1 Neural Avatar and Scene Representation

In this section, we formally introduce the geometry and appearance representations for humans and scenes in HSR. To account for dynamic human motion, we explain how skeletal deformation is used to model articulated human bodies.

Geometry Representation. We model the canonical human (H) and scene geometry (S) as two neural networks, f_{sdf}^H and f_{sdf}^S, where each one predicts the signed distance value for any 3D point in its respective space. Specifically:

$$f_{\text{sdf}}^H : \mathbb{R}^{3+n_\theta} \rightarrow \mathbb{R}^{1+n_z};\ (\mathbf{x}_c, \boldsymbol{\theta}) \mapsto (\xi^H, \mathbf{z}^H), \tag{1}$$

$$f_{\text{sdf}}^S : \mathbb{R}^3 \rightarrow \mathbb{R}^{1+n_z};\ \mathbf{x}_s \mapsto (\xi^S, \mathbf{z}^S). \tag{2}$$

The human shape f_{sdf}^H is modeled in the canonical space and deformed into the observation space with LBS deformations [41]. To capture pose-dependent

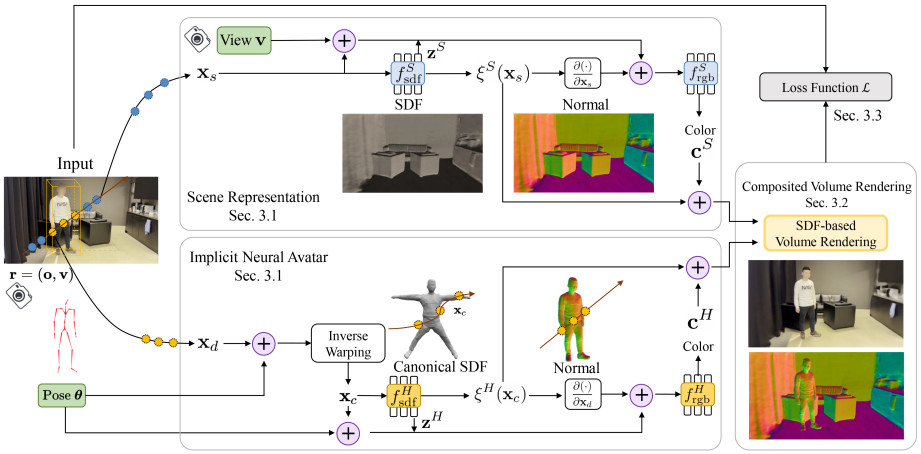

Fig. 2: Method overview. Given a ray, we sample two sets of points along it, one for the human $\{\mathbf{x}_d\}_{i=1}^N$ and one for the scene $\{\mathbf{x}_s\}_{i=1}^N$. The points for the human are sampled only inside the 3D human bounding box. We model the human H in a canonical space with an SDF-based shape neural network f_{sdf}^H [21] and a texture network f_{rgb}^H, then deform it using skinning techniques. Similarly, the scene is represented with neural fields f_{sdf}^S and f_{rgb}^S [74]. The outputs from both branches are used to composite a final image via SDF-based volume rendering. This allows us to jointly optimize the scene and human, treating the problem of 3D human-scene reconstruction holistically.

local non-rigid deformations such as dynamically changing wrinkles on clothes, we concatenate the human pose $\boldsymbol{\theta}$ as an additional input to $\mathbf{x}_c$. The pose parameters $\boldsymbol{\theta}$ are the SMPL [41] pose parameters with dimensionality n_θ. Each network outputs the signed distance value, ξ^L and a global geometry feature $\mathbf{z}^L$, where $L \in \{H, S\}$. The shape $\mathcal{S}^L$ is then given by the zero-level set of f_{sdf}^L:

$$\mathcal{S}^L = \{\ \mathbf{x} \mid f_{\text{sdf}}^L(\cdot) = 0\ \}. \tag{3}$$

We jointly optimize the shape and pose parameters during training, which naturally encourages accurate pixel-level alignment and smooth motion. Compared to the 3D spatial location, the pose parameters are relatively high-dimensional. This can lead to overfitting shapes to the pose parameters instead of achieving temporally consistent canonical shapes. To address this issue, we perform linear dimension reduction on the pose parameters, enable the pose condition at a later stage, and periodically reset the pose parameters to zero.

Appearance Representation. The appearance of the human and the scene is modeled via two neural networks f_{rgb}^L, which predict color values for 3D points:

$$f_{\text{rgb}}^H : \mathbb{R}^{3+n_\theta+n_z+3} \rightarrow \mathbb{R}^3;\ (\mathbf{x}_c, \boldsymbol{\theta}, \mathbf{z}^H, \mathbf{n}_d) \mapsto \mathbf{c}^H, \tag{4}$$

$$f_{\text{rgb}}^S : \mathbb{R}^{3+3+n_z+3} \rightarrow \mathbb{R}^3;\ (\mathbf{x}_s, \mathbf{v}, \mathbf{z}^S, \mathbf{n}_s) \mapsto \mathbf{c}^S. \tag{5}$$

Again, the human network f^H_{rgb} operates in the canonical space and is conditioned on the SMPL pose $\boldsymbol{\theta}$. In contrast, the scene network f^S_{rgb} is conditioned on the viewing direction $\mathbf{v}$. Both networks f^L_{rgb} receive the global geometry feature $\mathbf{z}^L$ computed by the shape network f^L_{sdf}. Lastly, to achieve better disentanglement of geometry and appearance, we also condition f^L_{rgb} on the respective normals :

$$\mathbf{n}_d = \frac{\partial \xi^H(\mathbf{x}_c, \boldsymbol{\theta})}{\partial \mathbf{x}_d}, \qquad \mathbf{n}_s = \frac{\partial \xi^S(\mathbf{x}_s)}{\partial \mathbf{x}_s}. \tag{6}$$

Skeletal Deformation. We model human networks in a canonical and pose-independent space. To map between the canonical and the posed space, we use skeletal deformations following [21]. Given the bone transformation matrix $\mathbf{B}_i$ for joint $i \in \{1, ..., n_b\}$ extracted from $\boldsymbol{\theta}$, we can map between a pair of the canonical point $\mathbf{x}_c$ and hte corresponding deformed point $\mathbf{x}_d$ via linear blend skinning (LBS) and its inverse:

$$\mathbf{x}_d = \sum_{i=1}^{n_b} w_c^i \mathbf{B}_i \, \mathbf{x}_c, \quad \mathbf{x}_c = (\sum_{i=1}^{n_b} w_d^i \mathbf{B}_i)^{-1} \, \mathbf{x}_d, \tag{7}$$

where n_b denotes the number of bones in the transformation, and $\mathbf{w}_{(\cdot)}$ represents the skinning weights for $\mathbf{x}_{(\cdot)}$. We choose w_d to be the skinning weight of the SMPL vertex that is closest to $\mathbf{x}_d$ (and analogously for w_c and $\mathbf{x}_c$ in the canonical space). With this deformation, the normals $\mathbf{n}_d$ can be computed as [78]:

$$\mathbf{n}_d = \frac{\partial \xi^H(\mathbf{x}_c, \boldsymbol{\theta})}{\partial \mathbf{x}_c} \frac{\partial \mathbf{x}_c}{\partial \mathbf{x}_d} = \frac{\partial \xi^H(\mathbf{x}_c, \boldsymbol{\theta})}{\partial \mathbf{x}_c} (\sum_{i=1}^{n_b} w_d^i \mathbf{B}_i)^{-1}. \tag{8}$$

Choice of Coordinate System. Previous works on human reconstruction [21, 33] position cameras based on their relative location to the canonical human body. In contrast to a human-centric coordinate system, we place everything within a consistent global coordinate system. This choice not only facilitates spatially coherent 3D human reconstructions but also improves human-scene decomposition through multi-view photometric consistency. As a result, this leads to the correct reconstruction of foreground occluders, which is essential to model occlusion. We demonstrate the effectiveness of such a choice of coordinate system in the supplementary material. To align the estimated human pose from each frame and the scene into a global coordinate system, we use the contact priors similar to NeuMan [34] to resolve the scale ambiguity. This gives us a rough estimation of the human scale in the scene. The scale can be further refined in the optimization stage to match the 2D projection of human bodies.

3.2 Compositional Volume Rendering

We normalize the scene shape field into a sphere with a pre-defined radius and align the deformed human shape with the scene by estimating a relative scale to ensure proper 2D projection and 3D contacts [34]. We apply SDF-based volume

rendering for both the dynamic human and static scenes. The final pixel value is then attained via compositing these two components.

Shape-aware Sampling. Given a ray $\mathbf{r} = (\mathbf{o}, \mathbf{v})$ with camera center $\mathbf{o}$ and ray direction $\mathbf{v}$, we sample N points for the human and scene networks (i.e., $2N$ in total) as $\mathbf{x}^i = \mathbf{o} + t^i\mathbf{v}$. We follow VolSDF [70] and use a two-stage sampling procedure with uniform and inverse CDF sampling. To better disentangle the human and the background even under strong occlusions, we design a shape-aware sampling strategy that leverages 3D body models (e.g., SMPL) to guide the sampling process. Specifically, we only sample points for the human shape and texture fields on the part of the ray that intersects with the axis-aligned 3D bounding box derived from the SMPL body estimation. The sampling range for the scene is determined by the camera center and the pre-defined sphere. We show in Fig. 5 that this sampling strategy is crucial to separate the human from the scene cleanly.

Surface-based Volume Rendering. For the surface-based volume rendering [70], we convert the SDF to a density value σ by applying the scaled Laplace distribution's Cumulative Distribution Function (CDF, $L \in \{H, S\}$ and $\beta, \gamma > 0$ are learnable parameters):

$$\sigma^L(\mathbf{x}) = \beta \left(\frac{1}{2} - \frac{1}{2} \operatorname{sign}(\xi^L(\mathbf{x}))(1 - \exp(-\frac{|\xi^L(\mathbf{x})|}{\gamma})) \right). \tag{9}$$

Scene Composition. To determine the final pixel value for a ray $\mathbf{r}$, we raycast the human and scene volumes separately. We then sort the sampled points based on the distances to the camera center, followed by a scene composition step for the volume integration:

$$C(\mathbf{r}) = \sum_{i=1}^{2N} \tau_i \mathbf{c}^L(\mathbf{x}^i), \tag{10}$$

$$\tau_i = \exp\left(-\sum_{j=1}^{i-1} \sigma^L(\mathbf{x}^j)\delta^j \right) \left(1 - \exp(-\sigma^L(\mathbf{x}^i)\delta^i)\right), \tag{11}$$

where $L = H$ if a point $\mathbf{x}^i$ is sampled for the human model (i.e., $\mathbf{x}^i \in \mathcal{H}$) and $L = S$ otherwise. δ^i is the distance between two adjacent samples. Here, the accumulated alpha value of either model for a pixel can be obtained by $\alpha^H(\mathbf{r}) = \sum_{\mathbf{x}_i \in \mathcal{H}} \tau_i$ and $\alpha^S(\mathbf{r}) = \sum_{\mathbf{x}_i \notin \mathcal{H}} \tau_i$. Similarly, we compute the depth $D(\mathbf{r})$ and normal $N(\mathbf{r})$ of the surface intersecting the current ray as:

$$D(\mathbf{r}) = \sum_{i=1}^{2N} \tau_i t^i, \; N(\mathbf{r}) = \sum_{\mathbf{x}_i \in \mathcal{H}} \tau_i \mathbf{n}_d(\mathbf{x}^i) + \sum_{\mathbf{x}_i \notin \mathcal{H}} \tau_i \mathbf{n}_s(\mathbf{x}^i). \tag{12}$$

3.3 Global Optimization

To train the human and scene models jointly from videos, we formulate the training as global optimization over all K frames with the following losses.

Image Reconstruction Loss. We calculate the L_1-distance between the rendered color $C(\mathbf{r})$ and the pixel's RGB value $\hat{C}(\mathbf{r})$ to compute the image reconstruction loss $\mathcal{L}^k_{\text{rgb}}$ for frame k:

$$\mathcal{L}^k_{\text{rgb}} = \frac{1}{|\mathcal{R}^k|} \sum_{\mathbf{r}\in\mathcal{R}^k} |C(\mathbf{r}) - \hat{C}(\mathbf{r})|, \tag{13}$$

where $\mathcal{R}^k$ denotes the sampled rays for frame k.

Interpenetration Loss. We penalize human-scene interpenetration of implicit surfaces using $\mathcal{L}^k_{\text{inter}}$. We select points that are predicted to be simultaneously inside the human and the scene by querying the individual SDF values in both shape networks, i.e., $\mathcal{S}^k = \{\mathbf{x}_c \colon \xi^H(\mathbf{x}_c) < 0\} \cap \{\mathbf{x}_c \colon \xi^S(LBS(\mathbf{x}_c)) < 0\}$, where LBS is the linear blend skinning function:

$$\mathcal{L}^k_{\text{inter}} = \frac{1}{|\mathcal{S}^k|} \sum_{\mathbf{x}_c\in\mathcal{S}^k} \xi^H(\mathbf{x}_c) \cdot \xi^S(LBS(\mathbf{x}_c)). \tag{14}$$

Mask Loss. We propose adding a foreground and background mask loss to help guide human-scene separation and refine human pose and geometry. Specifically, we apply a L_1 loss on the occlusion-aware foreground accumulated weights and background accumulated weights against the human mask $\mathcal{M}^k$:

$$\mathcal{L}^k_{\text{mask}} = \sum_{\mathbf{r}\in\mathcal{M}^k} |\alpha^H(\mathbf{r}) - 1| + \sum_{\mathbf{r}\notin\mathcal{M}^k} |\alpha^S(\mathbf{r}) - 1|. \tag{15}$$

Due to the inaccuracy in the generated human mask, we decrease the strength of mask supervision as training progresses. More details on mask generation and the effectiveness of such mask loss are provided in the supplementary material.

Self-Decomposition Loss. To mitigate incorrect mask supervision and let the model refine the human-scene separation, we use a scene decomposition loss [21]. Such a loss encourages the density distribution to be sharp and sparse so the network can figure out the accurate boundary. This loss includes two parts: an opacity regularization L_{sparse} to regularize the ray opacity via the canonical human shape and a BCE loss that penalizes deviations of the ray opacities from a binary $\{0, 1\}$ distribution:

$$\mathcal{L}^k_{\text{sparse}} = \frac{1}{|\mathcal{R}^k_{\text{off}}|} \sum_{\mathbf{r}\in\mathcal{R}^k_{\text{off}}} |\alpha^H(\mathbf{r})|, \tag{16}$$

$$\mathcal{L}^k_{\text{BCE}} = -\frac{1}{|\mathcal{R}^k|} \sum_{L\in\{H,S\}} \sum_{\mathbf{r}\in\mathcal{R}^k} \big(\alpha^L(\mathbf{r}) \log(\alpha^L(\mathbf{r})\big) + (1 - \alpha^L(\mathbf{r})) \log(1 - \alpha^L(\mathbf{r}))), \tag{17}$$

where the set of rays that do not intersect with the human is denoted as $\mathcal{R}^k_{\text{off}}$. We summarize both terms as $\mathcal{L}^k_{\text{dec}} = \lambda_{\text{BCE}}\mathcal{L}^k_{\text{BCE}} + \lambda_{\text{sparse}}\mathcal{L}^k_{\text{sparse}}$. We gradually increase the weight for this loss during the optimization.

Geometric Prior Loss. Using RGB images only to reconstruct complex and dynamic 3D scenes is an under-constrained problem as there exists an infinite number of photo-consistent explanations [4, 75]. This is even worse for the scene reconstruction due to the lack of explicit across-frame constraints. To this end, we follow [74] and leverage monocular geometric priors as additional supervision signals for the background reconstruction. Specifically, we use a pre-trained Omnidata model [14] to predict a depth map $\hat{D}$ and a normal map $\hat{N}$ for frame k. We then enforce the depth and normal consistency between the 2D estimations and the rendered ones:

$$\mathcal{L}^k_{\text{depth}} = \sum_{\mathbf{r}\in\mathcal{R}^k} \|(w\hat{D}(\mathbf{r}) + q) - D(\mathbf{r})\|^2, \tag{18}$$

$$\mathcal{L}^k_{\text{normal}} = \sum_{\mathbf{r}\in\mathcal{R}^k} \|\hat{N}(\mathbf{r}) - N(\mathbf{r})\|_1 + \|1 - \hat{N}(\mathbf{r})^\top N(\mathbf{r})\|_1, \tag{19}$$

where w and q are learnable scale and shift parameters. The final geometry prior loss is then given by $\mathcal{L}^k_{\text{geo}} = \lambda_{\text{depth}}\mathcal{L}^k_{\text{depth}} + \lambda_{\text{normal}}\mathcal{L}^k_{\text{normal}}$. More details on the depth loss are provided in the supplementary material.

Eikonal Loss. Like IGR [18], we force the shape networks f^H_{sdf}, f^S_{sdf} to satisfy the Eikonal equation:

$$\mathcal{L}^k_{\text{eik}} = \mathbb{E}_{\mathbf{x}_c}\left(\|\nabla f^H_{\text{sdf}}(\mathbf{x}_c)\| - 1\right)^2 + \mathbb{E}_{\mathbf{x}_s}\left(\|\nabla f^S_{\text{sdf}}(\mathbf{x}_s)\| - 1\right)^2. \tag{20}$$

Full Loss. Given a video sequence with K input frames, we minimize the combined loss function:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta}) = \sum_{k=1}^{K} &\mathcal{L}^k_{\text{rgb}}(\boldsymbol{\Theta}) + \lambda_{\text{inter}}\mathcal{L}^k_{\text{inter}}(\boldsymbol{\Theta}) + \lambda_{\text{mask}}\mathcal{L}^k_{\text{mask}}(\boldsymbol{\Theta}) + \\ &\lambda_{\text{dec}}\mathcal{L}^k_{\text{dec}}(\boldsymbol{\Theta}) + \lambda_{\text{geo}}\mathcal{L}^k_{\text{geo}}(\boldsymbol{\Theta}) + \lambda_{\text{eik}}\mathcal{L}^k_{\text{eik}}(\boldsymbol{\Theta}),\end{aligned} \tag{21}$$

where $\boldsymbol{\Theta}$ is the set of all optimizable parameters for the human and scene model.

4 Experiment

4.1 Evaluation Protocol

Synthetic Human Scene Dataset (SHSD). The currently available human motion datasets that include scene scans solely consist of SMPL or SMPL-X registrations, lacking precise ground truth for detailed human surface geometry [23, 26, 30]. To fill this gap, we introduce a new dataset named SHSD with ground-truth information for the surface geometry and appearance of human subjects and their surrounding environments. We employ a multiview capture stage to capture and reconstruct dynamic human subjects [10] in high quality. Subsequently, we render these realistic subjects into virtual environments with Blender [11], utilizing the virtual environments from ReplicaCAD [56] and

CIRCLE [3] datasets. By combining life-like human scans with realistic virtual environments, we create a dataset for further analysis and research of holistic 3D human-scene reconstruction. In total, SHSD consists of six sequences. For more details, please refer to the supplementary material.

Real Dataset. We record six monocular sequences using a mobile phone to demonstrate the effectiveness of our approach on real-world data. Each sequence captures a person engaging in various poses within an everyday scene.

Baselines. We compare our methods with the state-of-the-art human reconstruction methods SelfRecon [33] and Vid2Avatar [21]. We also compare our results with the holistic reconstruction methods PPR [67] and Total-Recon [53]. Total-Recon [53] focuses on reconstructing articulated objects and scenes from RGB-D input. Similarly, PPR [67] reconstructs articulated objects and scenes from monocular videos, employing physical simulation to determine relative scale and optimize pose. Therefore, PPR [67] and Total-Recon [53] do not specifically focus on modeling humans explicitly.

Evaluation Metrics. For view synthesis, we report PSNR, SSIM [61], and LPIPS [76] for image quality. In addition, we report the intersection over union (Mask IoU) between ground truth and predicted human masks for evaluation on human-scene decomposition. For the evaluation of human mesh reconstruction, we use volumetric IoU (3D IoU), Chamfer distance (CD, in cm), and normal consistency (NC) as metrics. For the quality of holistic reconstruction, we report the cosine similarity of normal maps (N-cos, in range $(0, 1)$) and the L_1 distance of the depth map error (D-ℓ_1, in cm).

4.2 Evaluation on Human Reconstruction

We present comparisons between our methods and Vid2Avatar [21] and SelfRecon [33] qualitatively in Fig. 3 and quantitatively in Tab. 1. Specifically, our method achieves better human-scene decomposition as shown in the first row in Fig. 3 and higher mask IoU in Tab. 1. In addition, our method can deal with severe occlusion, while Vid2Avatar and SelfRecon fail drastically, as shown in the second row in Fig. 3. This is because we place everything in a global consistent coordinate system, where obstructing objects can be reconstructed, and the spatial relationship between human and scene objects can be correctly inferred. In contrast, Vid2Avatar and SelfRecon hold the assumption that humans are not occluded and adopt a human-centric coordinate system which breaks the multiview consistency for the scene. In the third row in Fig. 3, we show that our method is more robust to initial pose estimation error as we optimize the pose parameters and the accurate 2D mask loss. More discussion on this is included in the supplementary material.

4.3 Evaluation on Holistic Reconstruction

We further compare our method with PPR [67] and Total-Recon [53] in Fig. 4 and Tab. 2. PPR utilizes RGB video input and leverages differentiable physics simulations to handle foot contact. However, in physics simulations, PPR only takes

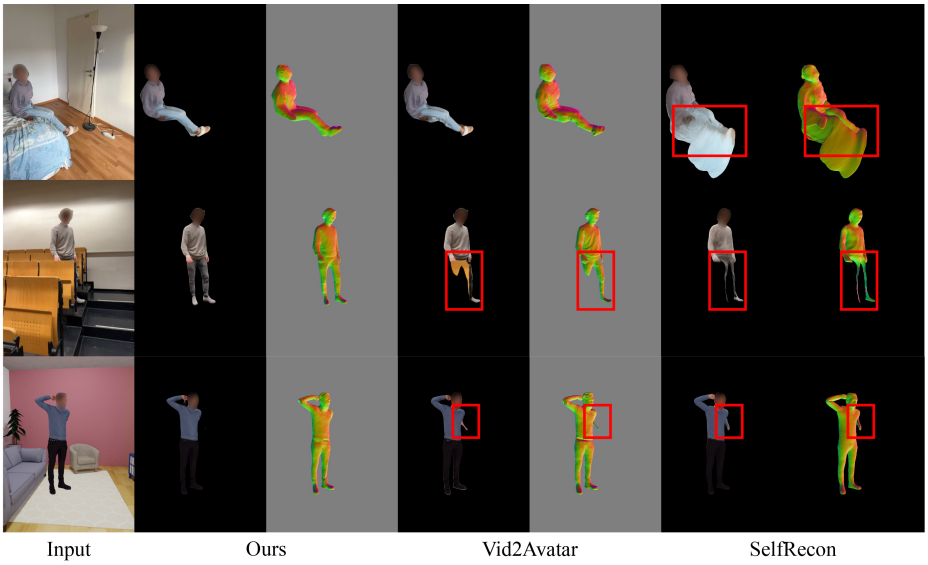

Fig. 3: Qualitative results on human reconstruction. Our method shows superior human-scene decomposition (first row), enhanced robustness to occlusion (second row), and improved handling of inaccurate poses (third row) compared to baselines.

	PSNR ↑	SSIM ↑	LPIPS ↓	Mask IoU ↑	3D IoU ↑	CD ↓	NC ↑
SelfRecon [33]	20.84	0.932	0.077	0.936	0.736	3.004	0.735
Vid2Avatar [21]	22.12	0.936	0.065	0.923	0.732	2.626	0.768
Ours	**22.34**	**0.937**	**0.057**	**0.954**	**0.757**	**2.349**	**0.782**

Table 1: Quantitative evaluation of human reconstruction on SHSD. Our method achieves better rendering quality, more accurate human-scene decomposition, and more faithful geometry reconstruction.

the ground as a plane into account, neglecting other objects in the scene (e.g., the bed, as shown in Fig. 4). Furthermore, PPR reconstructs the wall and the floor of the scene but struggles with detailed objects such as furniture, resulting in notable geometry errors. In contrast, our method considers the entire scene and makes better use of human priors, thus outperforming PPR significantly on human-scene reconstruction. Although Total-Recon takes depth maps as an additional input, the results presented in Fig. 4 demonstrate that our method performs much better even with only RGB signals. For quantitative comparison, PPR fails on certain sequences within our dataset, thus we only report the results on a subset of our dataset in Tab. 2. Regarding Total-Recon, since our dataset lacks simulated LIDAR inputs from a mobile device, we exclusively compare with Total-Recon on real data in Fig. 4. Please refer to our accompanying video and supplementary materials for more comparisons and results.

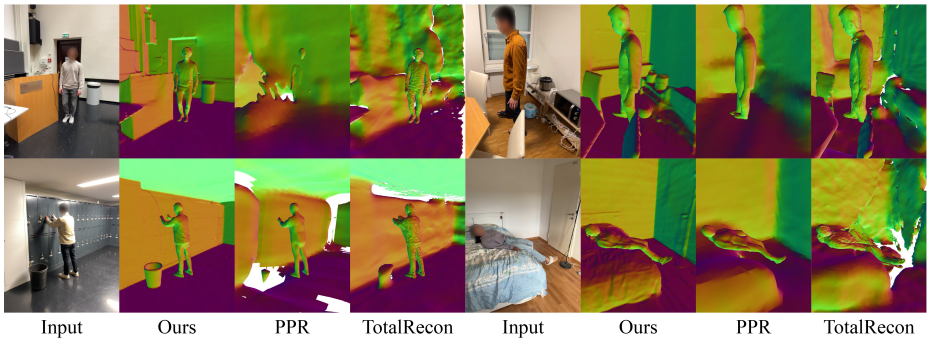

Fig. 4: Qualitative results on holistic reconstruction. Our method achieves much better reconstruction quality even without depth input as in TotalRecon [53].

	N-cos ↑	D-ℓ_1 ↓
PPR [67]	0.838	42.29
Ours	**0.936**	**15.14**

Table 2: Holistic reconstruction quality on SHSD. We report normal cosine on normal maps and depth ℓ_1 distance (cm) on depth maps. Our method achieves much better reconstruction quality compared to PPR [67].

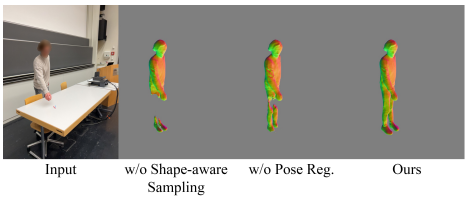

Fig. 5: Ablations for occlusion. Without shape-aware sampling or pose regularization, the human shape overfits to the observations, breaking temporal consistency across different frames.

4.4 Ablation Study

In this section, we present various ablation experiments to demonstrate the importance of HSR's design choices. The qualitative results in Fig. 5 illustrate techniques for a consistent human reconstruction, particularly under occlusion. Fig. 6 highlights the critical role of monocular geometric cues in monocular video-based scene reconstruction. Additional quantitative results are included in the supplementary material. Fig. 7 demonstrates the effectiveness of interpenetration loss in resolving human-scene interpenetration.

Consistent Human Reconstruction. When occlusion occurs, all supervision signals are utilized for the foreground occlusions, leaving the occluded human parts under-constrained. To achieve a reasonable human reconstruction in such a scenario, we need to rely on the temporal consistency of the canonical human space. However, the high dimensionality of pose parameters, compared to the three-dimensional location input in Eq. (1), causes the shape network to overfit to the pose condition, resulting in artifacts as depicted in Fig. 5. To address this issue, we propose to enable the pose condition only at a late stage to learn pose-dependent deformations for clothed humans. Additionally, we periodically

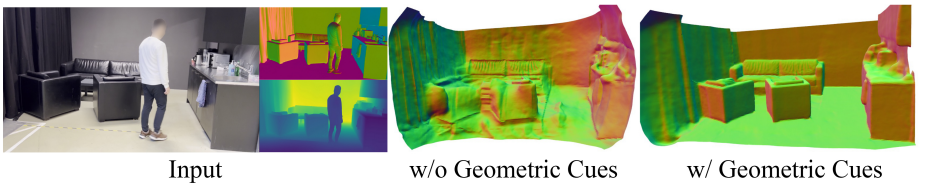

Fig. 6: Effect of monocular geometric cues. Depth and normal priors provide complementary supervision to RGB signals, resulting in much better reconstruction quality.

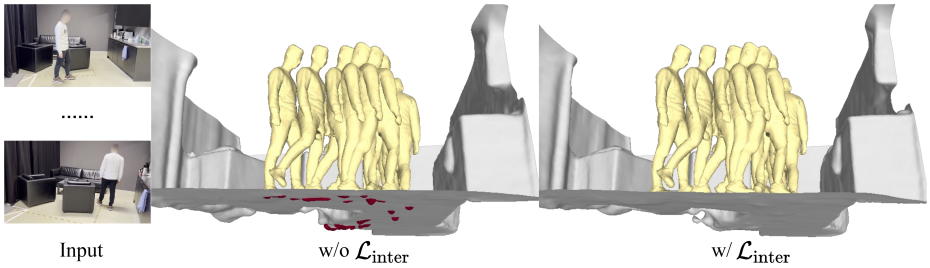

Fig. 7: Effect of the interpenetration loss. Vertices colored in red indicate the interpenetrated regions. Without $\mathcal{L}_{\text{inter}}$, the reconstructed human meshes unrealistically penetrates the floor.

set the pose condition for the shape network to zero, ensuring consistency with the canonical shape. This regularization technique helps achieve plausible human reconstruction even under strong occlusion.

Shape-aware Sampling. In Fig. 5, we demonstrate the importance of shape-aware sampling, which prevents artifacts such as truncated bodies. Without this strategy, the samples will be scattered along the entire ray due to the limitation of inverse CDF sampling. By applying a strong prior over a tight sampling range, all samples for the human networks are concentrated around the human body, minimally contributing to the final pixel color in case of occlusion. We include an illustration of this scenario in the supplementary material.

Monocular Geometric Cues. We use normal and depth maps predicted by Omnidata [14] to enhance our reconstruction quality. As shown in Fig. 6, these cues are particularly helpful for monocular reconstruction, especially in textureless regions such as walls and floors.

Human-Scene Interpenetration. Holistic reconstruction from images can result in surface interpenetration due to inaccuracies in pose estimation and geometry reconstruction. The proposed interpenetration loss mitigates collision around the contact regions as shown in Fig. 7. Our method effectively handles interactions with various scene elements through our general formulation of implicit scene modeling and the interpenetration loss. We provide a detailed examination of human-chair contacts in the supplementary material.

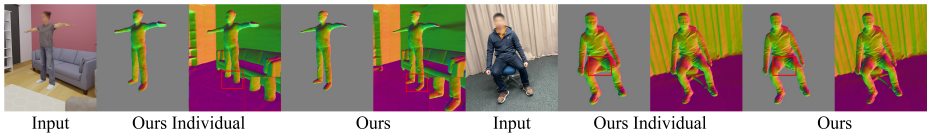

Fig. 8: Joint human-scene reconstruction. Training our method without a joint, global optimization (*Ours Individual*) leads to interpenetrations and inaccurate human-scene separation as highlighted.

Joint Human-Scene Reconstruction. To illustrate the necessity of joint human-scene reconstruction, we design a straightforward baseline method, combining MonoSDF [74] for scene reconstruction and Vid2Avatar [21] for human reconstruction. Specifically, we use human masks to separate the human from the scene and train each model independently. We refer to this baseline as *Ours Individual*. As shown in Fig. 8, *Ours Individual* suffers from spatial misalignment (left example) and segmentation errors (right example). In contrast, when optimizing the human model and the scene model simultaneously, our approach correctly determines the relative spatial order and effectively decouples the human and the scene.

5 Conclusion

We present HSR, a unified framework to jointly reconstruct static scenes with dynamic humans from monocular RGB videos. Our method models the human and the scene with neural implicit SDFs in a pose-independent space. Importantly, we optimize all parameters globally, making the human aware of the scene structure to better handle occlusions and interpenetration.
Limitations. In this work, we focus on the body – itself a difficult problem – and thus leverage SMPL for deformation. Most of our reconstructions contain blobby hands and blurred faces. To solve this, one needs to have accurate hand pose and facial expression estimation from relatively low-resolution observations. Integrating more expressive deformers such as SMPL-[H|X] for better hand and face reconstruction is an interesting future research direction.
Ethical Concerns. The collection of datasets, 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.

Acknowledgements

This work was partially supported by the Swiss SERI Consolidation Grant "AI-PERCEIVE". The authors thank all the participants of the captured datasets. Part of the computations were performed on the ETH Zürich Euler cluster.

References

1. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8387–8397 (2018)
2. Alldieck, T., Zanfir, M., Sminchisescu, C.: Photorealistic monocular 3d reconstruction of humans wearing clothing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
3. 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)
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. CVPR (2022)
5. Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2022)
6. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European conference on computer vision. pp. 561–578. Springer (2016)
7. Cao, Z., Gao, H., Mangalam, K., Cai, Q., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: European Conference on Computer Vision (ECCV) (2020)
8. Casado-Elvira, A., Comino Trinidad, M., Casas, D.: PERGAMO: Personalized 3d garments from monocular video. Computer Graphics Forum (Proc. of SCA), 2022 (2022)
9. Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L.E., Ramamoorthi, R., Su, H.: Deep stereo using adaptive thin volume representation with uncertainty awareness. In: CVPR (2020)
10. 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`
11. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018), `http://www.blender.org`
12. Dai, Y., Lin, Y., Wen, C., Shen, S., Xu, L., Yu, J., Ma, Y., Wang, C.: Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6792–6802 (June 2022)
13. Darmon, F., Bascle, B., Devaux, J.C., Monasse, P., Aubry, M.: Improving neural implicit surfaces geometry with patch warping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6260–6269 (2022)
14. 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)
15. Feng, Q., Liu, Y., Lai, Y.K., Yang, J., Li, K.: Fof: Learning fourier occupancy field for monocular real-time human reconstruction. In: NeurIPS (2022)

16. Fu, Q., Xu, Q., Ong, Y.S., Tao, W.: Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. arXiv preprint arXiv:2205.15848 (2022)
17. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: ICCV (2015)
18. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. arXiv preprint arXiv:2002.10099 (2020)
19. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2495–2504 (2020)
20. 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)
21. 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)
22. Guzov, V., Chibane, J., Marin, R., He, Y., Saracoglu, Y., Sattler, T., Pons-Moll, G.: Interaction replica: Tracking human–object interaction and scene changes from human motion. In: International Conference on 3D Vision (3DV) (March 2024)
23. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2021)
24. Habermann, M., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Deepcap: Monocular human performance capture using weak supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020)
25. Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Trans. Graph. **38**(2) (mar 2019). https://doi.org/10.1145/3311970, https://doi.org/10.1145/3311970
26. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: ICCV (2019)
27. He, T., Collomosse, J., Jin, H., Soatto, S.: Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 9276–9287. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/690f44c8c2b7ded579d01abe8fdb6110-Paper.pdf
28. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: Arch++: Animation-ready clothed human reconstruction revisited. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11046–11056 (October 2021)
29. Ho, H.I., Song, J., Hilliges, O.: Sith: Single-view textured human reconstruction with image-conditioned diffusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
30. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 13274–13285 (Jun 2022)

31. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. International Journal of Computer Vision (IJCV) (2024). https://doi.org/10.1007/s11263-024-01984-1
32. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: Arch: Animatable reconstruction of clothed humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3093–3102 (2020)
33. 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)
34. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: Neuman: Neural human radiance field from a single video. In: Proceedings of the European conference on computer vision (ECCV) (2022)
35. Jiang, Z., Guo, C., Kaufmann, M., Jiang, T., Valentin, J., Hilliges, O., Song, J.: Multiply: Reconstruction of multiple people from monocular video in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024)
36. Kaufmann, M., Song, J., Guo, C., Shen, K., Jiang, T., Tang, C., Zárate, J.J., Hilliges, O.: EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In: International Conference on Computer Vision (ICCV) (2023)
37. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: Proceedings International Conference on Computer Vision (ICCV). pp. 11127–11137. IEEE (Oct 2021)
38. Li, Z., Shimada, S., Schiele, B., Theobalt, C., Golyanik, V.: Mocapdeform: Monocular 3d human motion capture in deformable scenes. In: International Conference on 3D Vision (3DV) (2022)
39. Lin, W., Zheng, C., Yong, J.H., Xu, F.: Relightable and animatable neural avatars from videos. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
40. Long, X., Lin, C., Wang, P., Komura, T., Wang, W.: Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII. pp. 210–227. Springer (2022)
41. 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)
42. Ma, X., Gong, Y., Wang, Q., Huang, J., Chen, L., Yu, F.: Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In: ICCV. pp. 5732–5740 (2021)
43. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (sep 2018)
44. 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)
45. Monszpart, A., Guerrero, P., Ceylan, D., Yumer, E., Mitra, N.J.: imapper: interaction-guided scene mapping from monocular videos. ACM Transactions on Graphics (TOG) **38**(4), 1–15 (2019)
46. Moon, G., Nam, H., Shiratori, T., Lee, K.M.: 3d clothed human reconstruction in the wild. In: European Conference on Computer Vision (ECCV) (2022)

47. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9054–9063 (2021)
48. Ren, Y., Wang, F., Zhang, T., Pollefeys, M., Süsstrunk, S.: Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. arXiv preprint arXiv:2212.08067 (2022)
49. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2304–2314 (2019)
50. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 84–93 (2020)
51. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: Pigraphs: learning interaction snapshots from observations. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)
52. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
53. Song, C., Yang, G., Deng, K., Zhu, J.Y., Ramanan, D.: Total-recon: Deformable scene reconstruction for embodied view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17671–17682 (October 2023)
54. Su, S.Y., Bagautdinov, T., Rhodin, H.: Danbo: Disentangled articulated neural body representations via graph neural networks. In: European Conference on Computer Vision (2022)
55. Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In: Advances in Neural Information Processing Systems (2021)
56. 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)
57. Wang, F., Galliani, S., Vogel, C., Pollefeys, M.: Itermvs: Iterative probability estimation for efficient multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8606–8615 (2022)
58. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: Patchmatchnet: Learned multi-view patchmatch stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14194–14203 (2021)
59. Wang, J., Wang, P., Long, X., Theobalt, C., Komura, T., Liu, L., Wang, W.: Neuris: Neural reconstruction of indoor scenes using normal priors. In: 17th European Conference on Computer Vision. pp. 139–155. Springer (2022)
60. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. Advances in Neural Information Processing Systems **34**, 27171–27183 (2021)
61. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861

62. Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5610–5619 (2021)
63. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16210–16220 (June 2022)
64. Xiang, T., Sun, A., Wu, J., Adeli, E., Li, F.F.: Rendering humans from object-occluded monocular videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2023)
65. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: Implicit Clothed humans Obtained from Normals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13296–13306 (June 2022)
66. Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H.P., Theobalt, C.: Monoperfcap: Human performance capture from monocular video. ACM TOG **37**(2), 27:1–27:15 (May 2018). https://doi.org/10.1145/3181973, http://doi.acm.org/10.1145/3181973
67. Yang, G., Yang, S., Zhang, J.Z., Manchester, Z., Ramanan, D.: Ppr: Physically plausible reconstruction from monocular videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3914–3924 (October 2023)
68. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 767–783 (2018)
69. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent MVSNet for high-resolution multi-view stereo depth inference. In: CVPR (2019)
70. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021)
71. Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., Lipman, Y.: Multiview neural surface reconstruction by disentangling geometry and appearance. Advances in Neural Information Processing Systems **33** (2020)
72. Yi, H., Huang, C.H.P., Tzionas, D., Kocabas, M., Hassan, M., Tang, S., Thies, J., Black, M.J.: Human-aware object placement for visual environment reconstruction. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3949–3960 (2022). https://doi.org/10.1109/CVPR52688.2022.00393
73. Yi, H., Huang, C.H.P., Tzionas, D., Kocabas, M., Hassan, M., Tang, S., Thies, J., Black, M.J.: Human-aware object placement for visual environment reconstruction. In: Computer Vision and Pattern Recognition (CVPR) (Jun 2022)
74. 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)
75. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv preprint arXiv:2010.07492 (2020)
76. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
77. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In: European conference on computer vision (ECCV) (Oct 2022)

78. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: I M Avatar: Implicit morphable head avatars from videos. In: Computer Vision and Pattern Recognition (CVPR) (2022)
79. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)