

MultiPly: Reconstruction of Multiple People from Monocular Video in the Wild

Zeren Jiang^{*1} Chen Guo^{*1} Manuel Kaufmann¹ Tianjian Jiang¹

Julien Valentin² Otmar Hilliges¹ Jie Song¹

¹ETH Zürich ²Microsoft

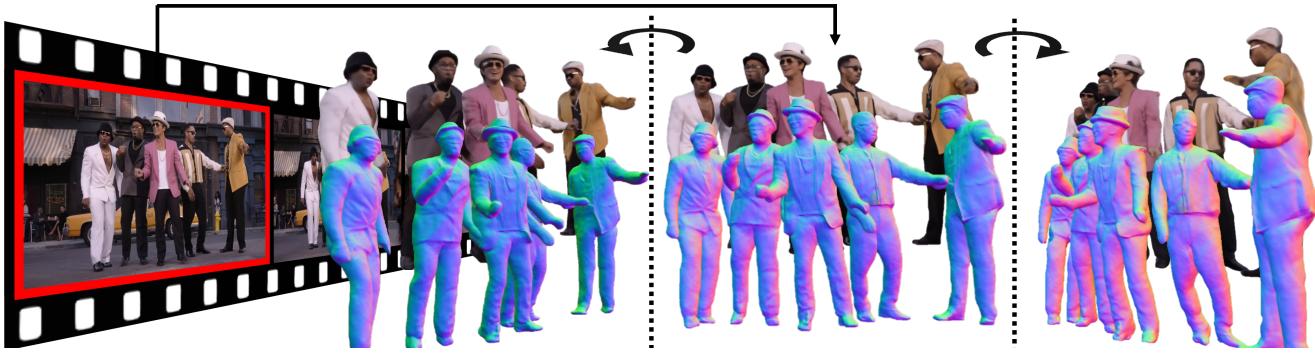


Figure 1. We propose MultiPly, a novel framework to reconstruct multiple people in 3D from in-the-wild monocular videos. Our method can recover the complete 3D human with high-fidelity shape and appearance, even in scenarios involving occlusions and close interactions.

Abstract

We present MultiPly, a novel framework to reconstruct multiple people in 3D from monocular in-the-wild videos. Reconstructing multiple individuals moving and interacting naturally from monocular in-the-wild videos poses a challenging task. Addressing it necessitates precise pixel-level disentanglement of individuals without any prior knowledge about the subjects. Moreover, it requires recovering intricate and complete 3D human shapes from short video sequences, intensifying the level of difficulty. To tackle these challenges, we first define a layered neural representation for the entire scene, composed by individual human and background models. We learn the layered neural representation from videos via our layer-wise differentiable volume rendering. This learning process is further enhanced by our hybrid instance segmentation approach which combines the self-supervised 3D segmentation and the promptable 2D segmentation module, yielding reliable instance segmentation supervision even under close human interaction. A confidence-guided optimization formulation is introduced to optimize the human poses and shape/appearance alternately. We incorporate effective objectives to refine human poses via photometric information and impose physically plausible constraints on human dynamics, leading to temporally consistent 3D reconstructions with high fidelity. The evaluation of our method shows the superiority over prior art on publicly available datasets and in-the-wild videos.

1. Introduction

Despite rapid progress in estimating the 3D shape from monocular videos of a single performer [11, 18, 35, 42], the analysis and reconstruction of several closely interacting people is still limited. This imbalance is unsatisfying, as group activities make up a significant portion of our lives. Although systems for multi-person reconstruction have been previously investigated, most require multi-view setups that constrain the capture area to a fixed volume and demand specialized equipment and expertise to operate [36, 47, 48]. Being able to reconstruct multiple people in detailed 3D geometry and appearance from monocular videos – which is also amenable to novice users – would facilitate many downstream tasks in AR/VR, such as the telepresence of groups of people, or the “replay” of social activities in 4D. Accomplishing this is fundamentally a challenging task since it requires accurate pixel-level disentanglement of individuals without a priori known geometries of the subjects. To make matters worse, the task is further complicated by depth ambiguities, complex human dynamics, and severe human-human occlusions – all of which have to be resolved from a single, short video clip.

In this paper, we introduce a novel method, called MultiPly, that provides a solution to this task: It takes a single video as input and outputs complete, high-quality, and separated 3D human geometry and appearance for individuals appearing in the scene (see Fig. 1). By embracing the promising paradigm of neural implicit functions for 3D representations, remarkable progress has recently been achieved in modelling detailed human geometry and ap-

^{*}These authors contributed equally to this work

pearance. While some methods require full supervision on 3D human scans that can be prohibitively expensive to acquire [35, 43], others only rely on readily available monocular videos [5, 11, 17, 20, 40] to fit articulated neural implicit fields with temporally consistent results. However, all of these methods are designed for only a single actor and thus neglect the complexities of the reconstruction task that arise from severe human-human occlusions. When applied to footage where multiple people engage in natural interactions, the aforementioned methods often result in corrupted or incomplete human reconstructions.

Our method builds on the promise of neural implicit fields, and presents a solution that overcomes the limitations of prior work. In doing so, several challenges must be addressed. First, 2D and 3D points of each subject must be precisely associated. Second, complete human models must be extracted and maintained from only a single video. Third, and most critically, the problem of person association and avatar creation are both significantly exacerbated by strong occlusions and imperfect human pose estimates.

To tackle these challenges, our approach is grounded in the following core concepts: i) We design a unified, temporally consistent representation of human shape and texture in canonical space that is applicable to all individuals. This design facilitates the integration of partial observations from the video sequence into a temporally coherent space, inherently maintaining a complete human body. ii) We establish a layered neural representation for the entire scene, wherein we parameterize the humans and the background as individual neural fields. The composition of these fields leads to a layered, interwoven representation that covers the entire space and which can be learned from a monocular RGB video via a proposed layer-wise differentiable volume rendering. iii) We introduce a hybrid instance segmentation approach that leverages the advantages of self-supervised scene decomposition in 3D and a learning-based promptable 2D segmentation module. This combination results in a robust and accurate separation of individuals. iv) To deal with imperfect pose estimates that might corrupt updates to the avatar model, we present a confidence-guided optimization formulation that alternately optimizes human poses and shape/appearance based on per-frame confidence measures. This way we incorporate effective objectives that refine the human poses through photometric information and impose physically plausible constraints on human dynamics.

In our experiments, we demonstrate that our framework leads to robust human instance segmentation and plausible pose estimates, achieving high-quality 3D reconstructions of multiple people even under extremely challenging visual complexities like severe occlusions (*cf.* Fig. 2). We meticulously ablate our method, which uncovers its essential components. Furthermore, we conduct comparisons with existing approaches in human reconstruction, novel view syn-

thesis, human instance segmentation, and pose estimation tasks, showing that our method outperforms prior art across various settings. In summary, our contributions are:

- a novel framework, MultiPly, to reconstruct multiple detailed 3D human models solely from in-the-wild monocular videos; and
- a robust instance segmentation approach that achieves a clean separation between people even under close interaction; and
- a confidence-guided optimization formulation that leads to temporally and spatially coherent 3D reconstructions of people with high fidelity.

2. Related Work

Monocular Single-Person Reconstruction Reconstructing an individual from monocular observations has emerged as a widely explored research challenge. Template-based approaches involve the tracking of a human template using 2D observations [12]. The assumption of a personalized template is unsuitable for more practical use cases. Follow-up works endeavor to remove this dependency by adding the vertex offsets on top of SMPL [1, 10]. Nevertheless, the explicit mesh representation is constrained by a fixed resolution and topology, incapable of representing fine-grained details. Learning-based methods that learn to regress 3D human shape from images have shown compelling results [2, 15, 34, 35, 42, 43, 52]. A major limitation of these methods is the necessity of high-quality 3D data for supervision and they fail to produce space-time coherent reconstructions over frames. Recent works employ neural rendering to train neural fields based on videos to obtain articulated human model [5, 6, 11, 17, 18, 20, 33, 37, 40]. Among these, Vid2Avatar [11] achieves compelling 3D reconstruction for a single subject but is not directly applicable to scenes with crowded people. Actually, none of the aforementioned methods can be directly deployed in more complicated multi-person scenarios. In contrast, we propose a novel framework that can faithfully reconstruct multiple people in the scene from a monocular video.

Monocular Multi-Person Reconstruction In contrast to the notable advancements in reconstructing the clothed human for an individual, limited emphasis has been placed on multi-person scenarios, which are evidently more applicable to our daily experiences. Most existing monocular works can only estimate the coarse body shapes of multiple people from monocular observations [4, 7, 14, 19, 21, 24, 25, 38, 39, 46]. Mustafa et al. [31] extend prior implicit methods to multiple people and recover spatially coherent 3D human shapes from an RGB image but mainly deal with cases where people are well-spaced and do not interact naturally in close range. Recently, more researchers have shifted the focus to multi-person scenarios [16, 36, 47, 51]. Even though these works achieve compelling instance-level

human reconstructions, they require expensive multi-view imaging systems. Concurrently, Cha et al. [3] propose to reconstruct multiple people from a single image. Such image-based methods usually fail to produce space-time coherent reconstructions over frames. Overall, monocular multi-person reconstruction is still an extremely under-explored problem. We propose MultiPly to take a significant stride towards addressing this formidable task.

Human Instance Segmentation Most works solve human or general object segmentation at the image level (i.e. 2D) [13, 22, 26, 41]. They are trained on images with human annotations to directly regress the segmentation masks during inference. More recently, a promptable segmentation model SAM has been developed to support flexible prompting along with input images [23]. However, SAM is a semantic segmentation method, which does not inherently support instance-level segmentation for humans. Therefore, meticulous prompts need to be designed for human instance segmentation tasks. Besides, these approaches are not able to produce sharp boundaries between individuals, especially for closely interacting people. More importantly, they do not always predict temporally coherent segmentation masks, as they focus on image-level segmentation only and incorporate no 3D knowledge. In this work, we optimize the instance segmentation masks on the fly by leveraging the promptability of SAM [23] and the self-supervised decomposition in 3D [11, 36].

3. Method

We present MultiPly, a novel framework for detailed geometry and appearance reconstruction of multiple people from in-the-wild monocular videos. The overview of our method is schematically illustrated in Fig. 2. Reconstructing multiple people in 3D from a short video without a priori known geometries is a challenging task due to complex human articulation, and strong occlusions. To tackle these challenges, we first define a unified, temporally consistent representation for humans and a layered neural representation for the entire scene (Sec. 3.1). The layered neural representation is then learned from images by performing our layer-wise differentiable volume rendering (Sec. 3.2). Given the self-supervised instance segmentation via occlusion-aware volume rendering, we further enhance the instance segmentation supervision by leveraging our evolving human surfaces in deformed space as progressively updated prompts for SAM which builds closed-loop refinement of instance segmentation in both 2D and 3D (Sec. 3.3). Finally, we formulate a confidence-guided optimization to alternately optimize human pose and shape/appearance (Sec. 3.4). We incorporate photometric information, robust instance segmentation supervision, and the inter-person objectives for pose refinement to achieve temporally and spatially coherent 3D reconstructions of people in high quality (Sec. 3.5).

3.1. Layered Neural Representation

Neural Avatars. For each human in the scene, we represent the 3D shape as an implicit signed-distance field (SDF) and the appearance as a texture field in canonical space, covering the entire space. When multiple people are present in the scene, it leads to a layered representation where the contributing SDFs are potentially interwoven. More specifically, we model the geometry and appearance of each person p in canonical space by a neural network f^p , which predicts the signed distance value s^p and the radiance value c^p at the query point \mathbf{x}_c^p in this space:

$$\mathbf{c}^p, s^p = f^p(\mathbf{x}_c^p, \boldsymbol{\theta}^p), \quad (1)$$

where $\boldsymbol{\theta}^p$ denotes the person’s pose parameters, which we concatenate to \mathbf{x}_c^p to model pose-dependent surface deformations. For simplicity, we use $f_c^p(\cdot)$ and $f_s^p(\cdot)$ to query \mathbf{c}^p and s^p from the network outputs.

Deformation Module. We follow a standard skeletal deformation based on SMPL [28] to find correspondences in canonical and deformed space. A canonical point \mathbf{x}_c^p is mapped to the deformed point \mathbf{x}_d^p via linear-blend skinning (LBS) based on SMPL transformation: $\mathbf{x}_d^p = T_{\text{smpl}}(\mathbf{x}_c^p, \boldsymbol{\theta}^p)$. Here, $T_{\text{smpl}}(\cdot)$ denotes the SMPL-based transformation derived from the body pose $\boldsymbol{\theta}^p$, which corresponds to LBS and is described in more detail in the Supp. Mat. Inversely, the canonical correspondence \mathbf{x}_c^p for point \mathbf{x}_d^p in deformed space is defined as $\mathbf{x}_c^p = T_{\text{smpl}}^{-1}(\mathbf{x}_d^p, \boldsymbol{\theta}^p)$. To invert LBS we use the SMPL skinning weight of the vertex closest to \mathbf{x}_d^p .

3.2. Layer-Wise Volume Rendering

We seek to reconstruct all human subjects in the scene. This requires different treatment compared to vanilla differentiable volume rendering that only works on a single static scene [30]. In contrast, on the basis of our layered neural avatar representation (Sec. 3.1), we introduce layer-wise volume rendering to handle dynamic scenes with multiple subjects and inter-occlusions. This is achieved by combining surface-based volume rendering [45] with the re-assembly of multiple human neural layers [16]. It is essential to note that the layer-wise volume rendering is inherently occlusion-aware.

Volume Rendering for Human Layers. For each sampled camera ray \mathbf{r} , we sample the points in the observation space along the ray based on the intersection between the oriented bounding box of the deformed SMPL model and the camera ray. Specifically, we sample N points $\{\mathbf{x}_{d,1}^p, \dots, \mathbf{x}_{d,N}^p\}$ inside the p -th intersected bounding box based on the two-stage sampling strategy proposed in [45]. Then, the occupancy o_i^p for the p -th person and the i -th sam-

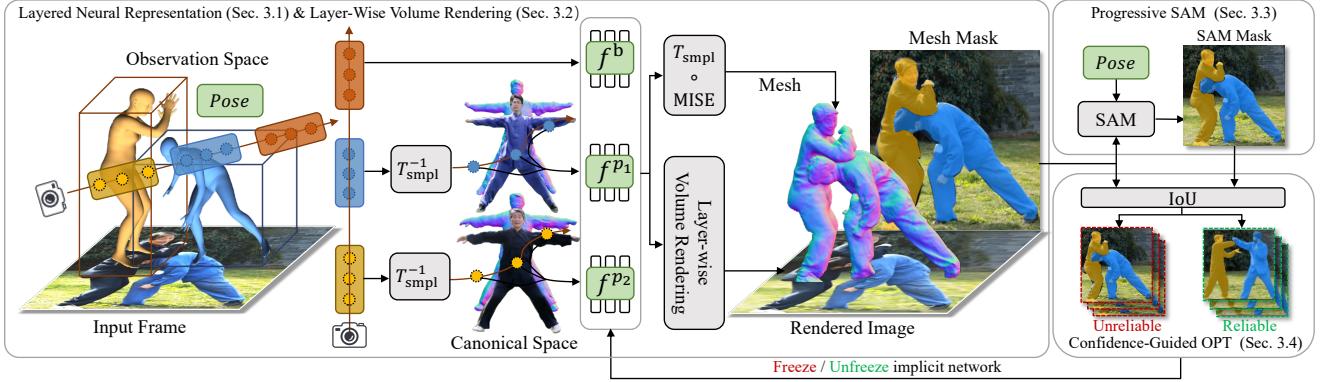


Figure 2. **Method overview.** Given an image and SMPL estimation, we sample human points along the camera ray based on the bounding boxes of SMPL bodies and the background points based on NeRF++. We warp sampled human points into canonical space via inverse warping and evaluate the person-specific implicit network to obtain the SDF and radiance values (Sec. 3.1). The layer-wise volume rendering is then applied to learn the implicit networks from images (Sec. 3.2). We build a closed-loop refinement for instance segmentation by dynamically updating prompts for SAM using evolving human models (Sec. 3.3). Finally, we formulate a confidence-guided optimization that only optimizes pose parameters for unreliable frames and jointly optimizes pose *and* implicit networks for reliable frames (Sec. 3.4).

pled point is calculated as follows:

$$\begin{aligned} o_i^p &= (1 - \exp(-\sigma_i^p \Delta x_i)), \\ \sigma_i^p &= \sigma \left(f_s^p \left(T_{\text{smpl}}^{-1} \left(\mathbf{x}_{d,i}^p, \theta^p \right), \theta^p \right) \right), \end{aligned} \quad (2)$$

where Δx_i is the distance between two adjacent sample points, and $\sigma(\cdot)$ is the scaled Laplace distribution’s Cumulative Distribution Function (CDF) defined in [45] to convert the signed distance s_i^p to volume density σ_i^p . Then we accumulate the radiance by performing numerical quadrature among the layered density field for multiple persons to obtain the color value:

$$\hat{C}^H = \sum_{i=1}^N \sum_{p=1}^P \left[o_i^p c_i^p \prod_{q=1}^P \prod_{j \in \mathcal{Z}_i^{q,p}} (1 - o_j^q) \right], \quad (3)$$

where P is the total number of subjects, \hat{C}^H is the rendered color of humans, and $\mathcal{Z}_i^{q,p}$ contains all indices of points (belonging to person q) whose depth value is lower than the depth value of the i -th point of person p , *i.e.* $\mathcal{Z}_i^{q,p} = \{j \in [1, N] \mid z(\mathbf{x}_{d,j}^q) < z(\mathbf{x}_{d,i}^p)\}$, where $z(\cdot)$ denotes the distance between the sampled point and the camera origin along the z-axis.

Scene Composition. We model the background in the same formulation as NeRF++ [49], denoted as f^b . We thus obtain a color value \hat{C}^B representing the background’s color, which is composited with \hat{C}^H via self-supervised decomposition following [11] to compute the final pixel color value \hat{C} . More details are shown in the Supp. Mat.

3.3. Progressive Prompt for SAM

Learning to disentangle and reconstruct multiple subjects by simply relying on the automatic separation through

layer-wise volume rendering is still a severely ill-posed problem. This is due to dynamically changing lighting effects (e.g., shadows) and potentially severe human-human occlusions and close contact. To this end, we propose to leverage the promptable segmentation model SAM [23] and design a progressive prompting strategy based on the evolving human models to provide robust instance segmentation supervision. In this section we describe how we design the prompt to get an updated SAM mask which we later use in the optimization (Sec. 3.4 and Sec. 3.5).

We define the p -th human shape to be the zero-level set of the signed distance function f_s^p and apply the *Multiresolution IsoSurface Extraction* (MISE) [29] to extract the mesh in canonical space, denoted as $\mathcal{S}_c^p = \langle \mathcal{V}_c^p, \mathcal{F}^p \rangle = \text{MISE}(f_s^p, \theta^p)$. Here, \mathcal{V}_c^p represents the extracted vertex set in canonical space, and \mathcal{F}^p denotes the corresponding faces. Then, the deformed vertex set in the observation space is:

$$\mathcal{V}_d^p = \{T_{\text{smpl}}(\mathbf{v}_c^p, \theta^p) \mid \mathbf{v}_c^p \in \mathcal{V}_c^p\}. \quad (4)$$

Similarly, the deformed mesh for the p -th person in the observation space is defined as $\mathcal{S}_d^p = \langle \mathcal{V}_d^p, \mathcal{F}^p \rangle$. Thus, we can obtain an instance mask $\mathcal{M}_{\text{mesh}}^p$ by differentiably rendering the deformed mesh. To improve efficiency, we opt for a differentiable rasterizer R rather than volume rendering:

$$\mathcal{M}_{\text{mesh}}^p = R(\mathcal{S}_d^p). \quad (5)$$

For the sake of clarity, we define $\mathcal{M} = 1$ to represent the inside and $\mathcal{M} = 0$ to indicate either the outside or occlusion by other meshes. The instance mask of deformed meshes $\mathcal{M}_{\text{mesh}}^p$ serves as one of the prompts for SAM refinement. We further provide points as input prompts. We begin by projecting the 3D keypoint candidates onto the image to obtain 2D keypoints $\mathcal{K}_{2d}^p = \{\Pi(\mathcal{J}(\theta^p, \beta^p))\}$, where

$\mathcal{J}(\theta^p, \beta^p)$ is the 3D SMPL keypoints given the pose θ^p and body shape β^p parameters for subject p , and Π is the camera projection function. The point prompts for the p -th subject are then defined by:

$$\begin{aligned}\mathcal{P}_+^p &= \{\mathbf{k} \in \mathcal{K}_{2d}^p \mid \mathcal{M}_{\text{mesh}}^p(\mathbf{k}) = 1\}, \\ \mathcal{P}_-^p &= \{\mathbf{k} \in \bigcup_{q \neq p} \mathcal{K}_{2d}^q \mid \mathcal{M}_{\text{mesh}}^p(\mathbf{k}) = 0\}.\end{aligned}\quad (6)$$

In other words, the positive point prompts \mathcal{P}_+^p include the 2D keypoints that are inside of the instance mask obtained from the deformed mesh and are outside of the instance masks of all other meshes. The negative point prompts are the union of all 2D keypoints of all other subjects that are outside of the projected mesh mask $\mathcal{M}_{\text{mesh}}^p$. The SAM instance mask $\mathcal{M}_{\text{sam}}^p$ is finally updated based on the combination of the mask and point prompts:

$$\mathcal{M}_{\text{sam}}^p = \text{SAM}(\mathcal{M}_{\text{mesh}}^p, \mathcal{P}_+^p, \mathcal{P}_-^p), \quad (7)$$

Note that $\mathcal{M}_{\text{sam}}^p$ are progressively updated during training.

3.4. Confidence-Guided Alternating Optimization

Human-human occlusions often lead to inaccurate pose and wrong depth order estimation. A naïve joint optimization for both the pose and shape parameters may end up with a suboptimal solution. To mitigate this, we introduce a confidence-guided optimization strategy to alternately optimize the human poses and shapes.

To avoid damaging shape updates that are due to wrong poses, we only optimize pose parameters for unreliable frames and jointly optimize pose *and* shape parameters for reliable frames. We treat the IoU between the projected mesh mask $\mathcal{M}_{\text{mesh}}^{p,i}$ and the refined SAM mask $\mathcal{M}_{\text{sam}}^{p,i}$ as our confidence measure for the p -th subject in frame i . We define reliable frames \mathcal{I}_r to be those frames with reliable poses based on the average IoU over all subjects:

$$\mathcal{I}_r = \left\{ \mathcal{I}_i \in \mathcal{I} \mid \frac{1}{P} \sum_{p=1}^P \text{IoU}(\mathcal{M}_{\text{mesh}}^{p,i}, \mathcal{M}_{\text{sam}}^{p,i}) \geq \alpha \right\}, \quad (8)$$

where \mathcal{I} are all frames, and $\mathcal{I} \setminus \mathcal{I}_r$ are unreliable frames. α is a confidence threshold which is set to be the median of all IoU values over the entire sequence. It's important to note that the confidence threshold α is dynamically updated during training and eventually inaccurate pose estimates are corrected and all frames will be used for joint optimization.

3.5. Objectives

Reconstruction Loss. We calculate the L_1 -distance between the rendered color $\hat{C}(\mathbf{r})$ and the image pixel's RGB value $C(\mathbf{r})$ over all sampled rays \mathcal{R} :

$$L_{\text{rgb}} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} |C(\mathbf{r}) - \hat{C}(\mathbf{r})|. \quad (9)$$

Instance Mask Loss. We first modify Eq. 3 to differentiably render the opacity $\hat{O}^p(\mathbf{r})$ per person per pixel:

$$\hat{O}^p(\mathbf{r}) = \sum_{i=1}^N \left[o_i^p \prod_{q=1}^P \prod_{j \in \mathcal{Z}_i^{q,p}} (1 - o_j^q) \right]. \quad (10)$$

Then the instance mask loss is calculated between the refined instance mask and the rendered pixel-wise opacity:

$$L_{\text{mask}} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} \sum_{p=1}^P |\mathcal{M}_{\text{sam}}^p(\mathbf{r}) - \hat{O}^p(\mathbf{r})|. \quad (11)$$

Eikonal Loss. Following [9], we sample points in the canonical space for each subject and enforce the Eikonal constraint to ensure f_s^p is a valid SDF:

$$L_e = \sum_{p=1}^P \mathbb{E}_{\mathbf{x}_c} (\|\nabla f_s^p(\mathbf{x}_c^p, \theta^p)\| - 1)^2. \quad (12)$$

We further introduce two inter-person objectives for pose refinement to ensure spatially coherent and physically plausible reconstructions. We apply these constraints explicitly on the deformed mesh to refine human poses while the deformed meshes \mathcal{S}_d^p are updated on the fly during training. To be specific, the following two additional losses are used periodically during training to optimize pose only:

Depth Order Loss. The effect of wrong depth order on the reconstruction quality can be severe, resulting in reversed geometry and texture. Similar to [19], we apply a depth order loss as follows:

$$L_{\text{depth}} = \sum_{(\mathbf{u}, p, q) \in \mathcal{D}} \log(1 + \exp(D_p(\mathbf{u}) - D_q(\mathbf{u}))), \quad (13)$$

where $\mathcal{D} = \{(\mathbf{u}, p, q) \mid p \neq q, \mathcal{M}_{\text{sam}}^p(\mathbf{u}) = \mathcal{M}_{\text{mesh}}^q(\mathbf{u}) = 1\}$ represents the set of pixels \mathbf{u} where we have depth ordering mistakes between the p -th and q -th persons. $D_p(\mathbf{u})$ denotes the depth of the p -th mesh for pixel \mathbf{u} .

Interpenetration Loss. We shoot a ray for the sampled vertex in \mathcal{V}_d^p in Eq. 4 to check the number of intersection with other meshes. Then, we use the parity of the number of intersections to determine whether that point is inside other meshes. Following [8], the interpenetration loss is calculated as follows:

$$L_{\text{inter}} = \sum_{p=1}^P \sum_{q=1, q \neq p}^P \|\mathcal{V}_{\text{in}}^{p,q} - \text{NN}(\mathcal{V}_{\text{in}}^{p,q}, \mathcal{S}_d^q)\|_2, \quad (14)$$

where $\mathcal{V}_{\text{in}}^{p,q}$ denotes the p -th person's vertex which is inside the q -th person's mesh, and $\text{NN}(\mathcal{V}, \mathcal{S})$ finds the nearest vertex in \mathcal{S} for each point in \mathcal{V} . Different from [8, 19] where they deploy the depth order and interpenetration loss on the naked parametric human model, we apply those two losses on our learned pixel-wise aligned clothed human meshes, leading to a more fine-grained optimization.

See Supp. Mat for more details about the final loss.

Metrics	Pose Estimation				Human Reconstruction			
	MPJPE \downarrow	MVE \downarrow	CD \downarrow	PCDR \uparrow	V-IoU \uparrow	C $- \ell_2 \downarrow$	P2S \downarrow	NC \uparrow
Initial pose	75.3	90.8	235.6	0.566	-	-	-	-
Layer-wise volume rendering	71.5	86.2	245.6	0.608	0.773	3.14	2.87	0.752
+ Progressive SAM	71.2	85.9	246.2	0.609	0.786	2.68	2.42	0.784
+ Confidence-guided OPT	69.4	83.6	218.4	0.709	0.816	2.53	2.34	0.789

Table 1. **Quantitative ablation studies on Hi4D.** We demonstrate the importance of the proposed progressive prompt for SAM and confidence-guided alternating optimization. Both key components effectively contribute to the final reconstruction quality (cf. Fig. 3).

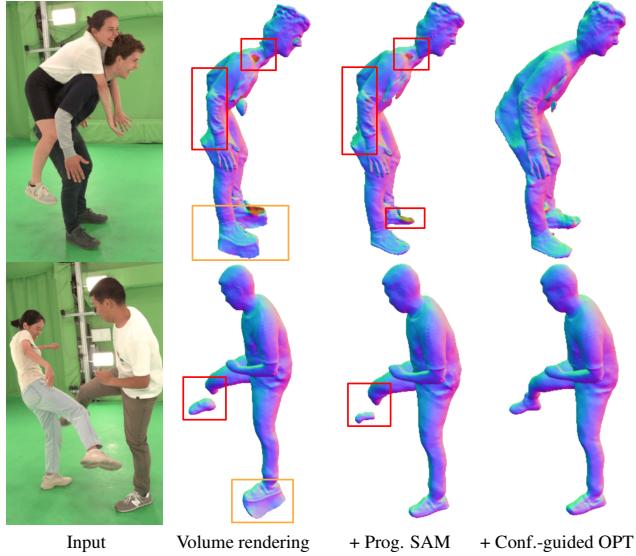


Figure 3. **Qualitative ablation studies.** Our progressive prompting strategy provides robust instance segmentation supervision and eliminates the noises caused by the environmental dynamic effects. The confidence-guided optimization further improves the reconstruction results and maintains complete human bodies.

4. Experiments

We first introduce the datasets and metrics used for evaluation. Next, ablation studies are conducted to demonstrate the effectiveness of our design choices. We then compare our proposed method with state-of-the-art approaches in four tasks, including human reconstruction, novel view synthesis, human instance segmentation, and pose estimation.

4.1. Datasets and Metrics

Hi4D [47]. This dataset contains challenging human interactions between pairs of people with ground truth meshes, human poses, and instance segmentation masks. We use Hi4D to evaluate our approach to all tasks.

Monocular Multi-huMan (MMM). Since the Hi4D dataset only contains two-person interactions with the static camera in the stage. In order to evaluate the generalization of our method, we collect a dataset called *Monocular Multi-huMan (MMM)* by using a hand-held smartphone, which contains six sequences with two to four persons in each sequence. Half of the sequences are captured in the stage with

ground truth annotations for quantitative evaluation and the others are captured in the wild for qualitative evaluation.

Metrics. We consider the following metrics for human mesh reconstruction evaluation: volumetric IoU (V-IoU), Chamfer distance ($C - \ell_2$) [cm], point-to-surface distance (P2S) [cm], and normal consistency (NC). Rendering quality is measured via three metrics: PSNR, SSIM, and LPIPS. We assess human pose estimation using four metrics: MPJPE [mm], MVE [mm], Contact Distance (CD) [mm], and Percentage of Correct Depth Relations (PCDR) with a threshold of 0.15m. Lastly, we report IoU, Recall, and F1 score for human instance segmentation.

4.2. Ablation study

As depicted in Tab. 1, ablation studies are conducted to demonstrate the effectiveness of the proposed progressive SAM prompt and confidence-guided optimization strategy. Both pose estimation and human reconstruction tasks are evaluated here. The initial human poses are obtained from TRACE [39] and ViTPose [44], more details are provided in the Supp. Mat. We initiate ablation studies based on our layer-wise volume rendering by naively optimizing the human pose and shape/appearance jointly without using SAM.

Human Reconstruction. Applying instance mask supervision based on progressively refined SAM outputs significantly improves the output quality and drastically reduces the reconstruction error (Chamfer distance) as quantitatively shown in Tab. 1. This is also confirmed by the qualitative results within the orange bounding boxes highlighted in Fig. 3, where layer-wise volume rendering purely relying on self-supervised segmentation fails to separate the dynamic shadows from the human, leading to noisy reconstructions. The reconstruction quality is further improved with the proposed confidence-guided optimization, as quantitatively indicated by the last row of Tab. 1. The red bounding boxes in Fig. 3 serve as visual evidence for the importance of our confidence-guided optimization. The presence of incomplete human bodies, such as the broken leg, disappeared back, and shrunken neck, is attributed to the incorrect depth order and pose estimation error. By temporarily freezing the implicit network for frames with unreliable poses, we circumvent such detrimental shape updates, leading to a complete human reconstruction.

Dataset	Method	V-IoU \uparrow	$C - \ell_2 \downarrow$	P2S \downarrow	NC \uparrow
Hi4D	ECON [43]	0.787	3.72	3.59	0.746
	V2A [11]	0.783	3.02	2.46	0.775
	Ours	0.816	2.53	2.34	0.789
MMM	ECON [43]	0.760	4.17	3.71	0.705
	V2A [11]	0.812	3.34	2.68	0.735
	Ours	0.826	2.89	2.40	0.757

Table 2. **Quantitative reconstruction evaluation.** Our method consistently outperforms all baselines on both datasets and all evaluation metrics (*cf.* Fig. 4).

Pose Estimation. We observe that this performance aligns with geometric reconstruction. Compared with the initial pose, we achieve not only more accurate individual poses (MPJPE and MVE) but also a better spatial arrangement between people, reflected on CD and PCDR.

4.3. Reconstruction Comparisons

To the best of our knowledge, there are few video-based reconstruction methods designed for clothed multiple people. Hence, we adapt two state-of-the-art reconstruction approaches to our setting for comparison. ECON [43] is an image-based 3D human reconstruction method capable of handling multi-person scenarios. While evaluating ECON, we discard frames with incorrect bounding-box detections for a fair comparison. Vid2Avatar (V2A) [11] is a video-based human reconstruction method designed for a single person. We extend V2A to multi-person scenarios by training a distinct model for each subject individually. Our method outperforms [11, 43] by a substantial margin on both datasets and all metrics (*cf.* Tab. 2). This disparity becomes more visible in qualitative comparisons shown in Fig. 4. When people closely interact, both ECON and V2A fail to recover complete human bodies but only output corrupted reconstructions (*e.g.*, missing legs/heads). Furthermore, they struggle with the initial depth order/pose error and produce spatially incorrect reconstructions in 3D. V2A tends to model environmental dynamic effects (*e.g.*, shadows) as the human body, resulting in noisy reconstructions. These artifacts are highlighted within the colored bounding boxes of Fig. 4. In contrast, our method generates complete human shapes with sharp boundaries and spatially coherent 3D reconstructions. We attribute this superiority to our proposed representation design and learning schemes.

4.4. Novel View Synthesis Comparisons

To the best of our knowledge, there are few novel view synthesis approaches particularly designed for clothed multiple people from monocular video. Hence, we adapt Shuai et al. [36], which is a state-of-the-art multi-person novel view synthesis approach from multi-view videos, to the monocular setting for a fair comparison. We share the same human pose initialization for training. Then, we use the ground

Method	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow
Shuai et al. [36]	0.898	19.6	0.1099
Ours	0.915	20.7	0.0798

Table 3. **Quantitative rendering evaluation on Hi4D.** Our method achieves better rendering quality (*cf.* Fig. 5).

Method	IoU \uparrow	Recall \uparrow	F1 \uparrow
SCHP [26]	0.937	0.983	0.982
Ours (Init.)	0.943	0.975	0.984
Ours (Progressive)	0.963	0.985	0.990

Table 4. **Quantitative instance segmentation evaluation on Hi4D.** Our method achieves the best segmentation accuracy.

truth human poses and camera parameters to render the novel view. As shown in Tab. 3 our method outperforms [36] on all metrics. Fig. 5 shows that the rendered image from [36] is more blurry and has noisy artifacts compared to ours. The reasons are twofold: 1) it lacks a reliable pose correction mechanism, leading to large inconsistency between human pose and image information during training, and 2) weakly-supervised decomposition cannot ensure robust instance segmentation under close human interaction. Our framework generates more plausible renderings with clearly sharp boundaries.

4.5. Instance Segmentation Comparisons

We compare our instance segmentation result from SAM after convergence with pretrained human instance segmentation network SCHP [26]. Tab. 4 reveals that our initial SAM outputs achieve comparable results with SCHP. However, the initial SAM masks are unsatisfactory due to the noisy prompt from the inaccurate SMPL estimation, leading to an incomplete and implausible reconstruction result (*e.g.*, the missing body part and self-interpenetration in the red bounding boxes) as shown in Fig. 6. Finally, our progressive prompting strategy based on our evolving human models helps to achieve temporally consistent and complete segmentation masks and high-quality reconstructions, surpassing the baseline methods.

4.6. Pose Estimation Comparisons

We conduct a comparison of our method with state-of-the-art bottom-up (TRACE [39]) and top-down (CLIFF [27]) multi-person pose estimation approaches. To adapt CLIFF for close human interaction, we employ ByteTrack [50] and linear interpolation to estimate missing persons caused by detector errors. Our approach consistently outperforms other baseline methods on all metrics as shown in Tab. 5. Specifically, our method shows its superiority in pose estimation accuracy of individuals (MPJPE and MVE) and more reasonable spatial arrangement between pairs of peo-

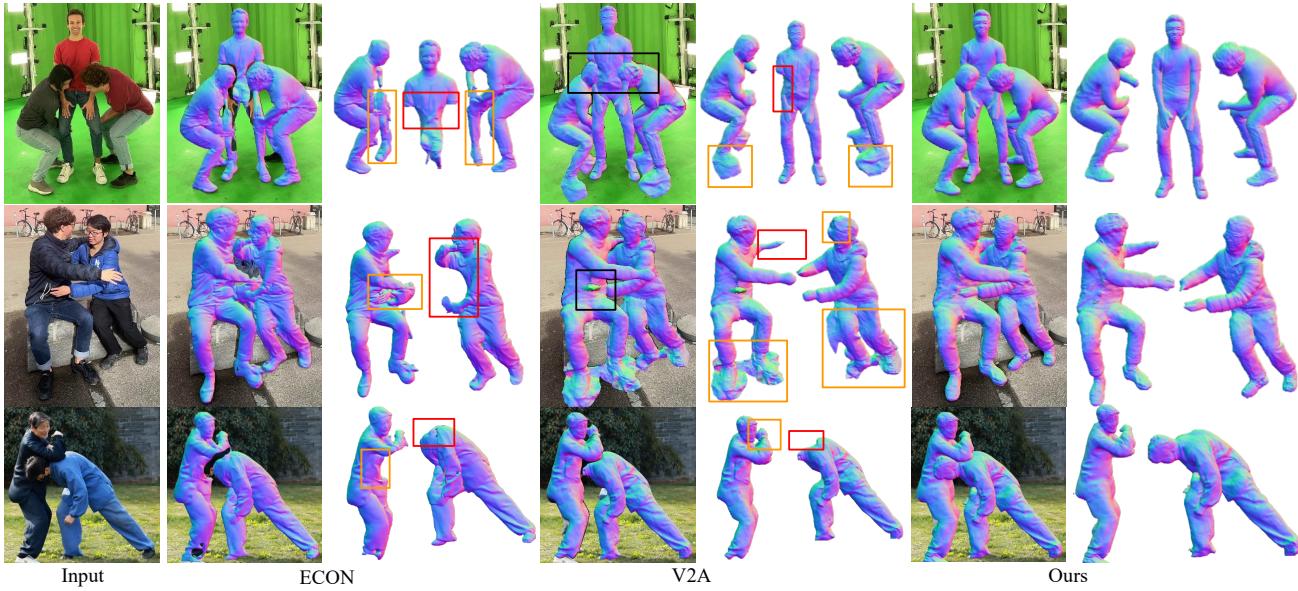


Figure 4. **Qualitative reconstruction comparison.** We show both the overlaid and separated reconstruction results for each method. **Red** bounding boxes: the incomplete reconstruction of the occluded part. **Orange** bounding boxes: incorrect instance segmentation results caused by the surrounding visual complexities. **Black** bounding boxes: inaccurate spatial arrangement due to pose error.

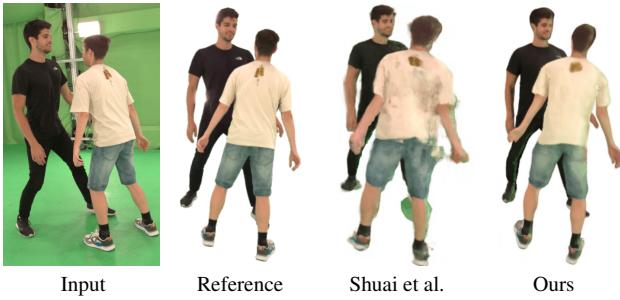


Figure 5. **Qualitative rendering comparison.** Our method achieves more plausible renderings with sharp boundaries.

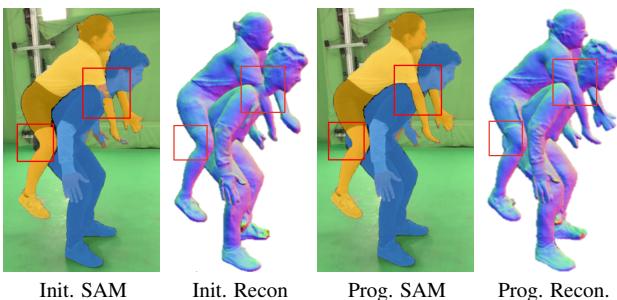


Figure 6. **Qualitative instance segmentation comparison.** Progressive prompting strategy provides more robust and fine-grained instance segmentation supervision compared to the initial SAM outputs, leading to higher quality of reconstructions.

ple (CD and PCDR). This is also confirmed by qualitative results. Please refer to the Supp. Mat.

Method	MPJPE \downarrow	MVE \downarrow	CD \downarrow	PCDR \uparrow
CLIFF [27]	85.7	102.1	351.7	0.606
TRACE [39]	95.6	115.7	249.4	0.603
Ours	69.4	83.6	218.4	0.709

Table 5. **Quantitative pose estimation evaluation on Hi4D.** Our method outperforms state-of-the-art multi-person pose estimation methods on all evaluation metrics.

5. Conclusion

In this paper, we present MultiPly, which for the first time produces temporally and spatially coherent 3D reconstructions of multiple people with high fidelity from monocular in-the-wild videos. We utilize carefully designed layered neural representation and dynamically refined instance segmentation supervision. We further introduce a confidence-guided optimization to learn human neural layers via layer-wise volume rendering. Our method is able to reconstruct multiple high-quality 3D human models in challenging scenarios involving close human interactions and strong inter-person occlusions.

Limitations: The complexity of our model increases linearly with the number of involved persons, making it inefficient for crowds. Our method does not explicitly model hands and we see the integration of an expressive human model [32] as a future direction. We discuss more limitations and potential negative societal impact in Supp. Mat.

Acknowledgement: This work was partially supported by the Swiss SERI Consolidation Grant ‘‘AI-PERCEIVE’’.

References

- [1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2
- [2] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [3] Junuk Cha, Hansol Lee, Jaewon Kim, Nhat Nguyen Bao Truong, Jae Shin Yoon, and Seungryul Baek. 3d reconstruction of interacting multi-person in clothing from a single image, 2024. 3
- [4] Yu Cheng, Bo Wang, and Robby T. Tan. Dual networks based 3d multi-person pose estimation from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1636–1651, 2023. 2
- [5] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 2
- [6] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. *arXiv*, 2023. 2
- [7] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 2
- [8] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazonvan, Vlad Olaru, and Cristian Sminchisescu. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. In *Advances in Neural Information Processing Systems*, pages 19385–19397. Curran Associates, Inc., 2021. 5
- [9] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5
- [10] Chen Guo, Xu Chen, Jie Song, and Otmar Hilliges. Human performance capture from monocular video in the wild. In *2021 International Conference on 3D Vision (3DV)*, pages 889–898. IEEE, 2021. 2
- [11] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. 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 (CVPR)*, 2023. 1, 2, 3, 4, 7
- [12] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3
- [14] Buzhen Huang, Jingyi Ju, Zhihao Li, and Yangang Wang. Reconstructing groups of people with hypergraph relational reasoning, 2023. 2
- [15] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [16] Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. In *ACM SIGGRAPH*, 2021. 2, 3
- [17] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [18] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2
- [19] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020. 2, 5
- [20] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2
- [21] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery, 2022. 2
- [22] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 3
- [23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 3, 4
- [24] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2
- [25] Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery, 2023. 2
- [26] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3, 7
- [27] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 7, 8
- [28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-

- person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3
- [29] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 4
- [30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 3
- [31] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. Multi-person implicit reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14474–14483, 2021. 2
- [32] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 8
- [33] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 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*, pages 9054–9063, 2021. 2
- [34] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morigima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2
- [35] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 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*, pages 84–93, 2020. 1, 2
- [36] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH Conference Proceedings*, 2022. 1, 2, 3, 7
- [37] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems*, 2021. 2
- [38] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [39] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6, 7, 8
- [40] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2
- [41] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 3
- [42] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 1, 2
- [43] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 7
- [44] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 6
- [45] Lior Yariv, Jitao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 3, 4
- [46] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [47] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 6
- [48] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yan-shun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40(4):1–18, 2021. 1
- [49] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 4
- [50] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. 2022. 7
- [51] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *IEEE Conference on Computer Vision (ICCV 2021)*, 2021. 2
- [52] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2