MV2MP: Segmentation Free Performance Capture of Humans in Direct Physical Contact from Sparse Multi-Cam Setups

Sergei Eliseev[1], Leonid Shtanko[2], Rasim Akhunzianov[2], Yaroslav Romanenko[2], and Anatoly Starostin[1]

[1] Yango Israel, Tel-Aviv, Israel
[2] Yandex, Moscow, Russia
{serg-turbo, leonsht, rakhunzy, yromko, starost}@yandex-team.ru

Abstract. This paper introduces a novel and robust approach for the performance capture of multiple humans engaged in direct physical interactions with very sparse RGB camera setups. Unlike existing methods that only perform well under specific conditions, such as when humans are relatively distant from each other, when a scene is surrounded by a large array of cameras, or when precise segmentation is available, our method operates without any of these requirements. We introduce a novel layered network architecture to represent the foreground and background together, as well as a tailored compositional volumetric rendering technique and objective functions, along with a new sampling method. These innovations enable the accurate reconstruction of humans engaged in direct physical interactions using only images and roughly estimated SMPL models. Our work demonstrates that our method is able not only to extract high-quality geometry of interacting people but also to provide segmentation and free viewpoint video, outperforming competitors that work in similar setups. Also we show the ability to improve the quality of the roughly estimated SMPL models. We have conducted experiments on a variety of scenes using the HI4D and CMU Panoptic datasets. The code and examples are available at `https://github.com/mv2mp/MV2MP`.

Keywords: Performance capture · Compositional Volumetric Rendering · Layered Network Architecture · Segmentation

1 Introduction

Human performance capture has gained a lot of attention in both academic and industry communities. Many content creators are especially interested in systems capable of capturing the performance of multiple people in close interactions. For example, the reconstruction of sports scenes demands accurate capture of humans performing actions not only in close proximity but in direct physical contact. Unlike static scenes, reconstructing high-fidelity dynamic human models faces inevitable challenges such as nonrigid motion, severe occlusions, deformations due to direct physical contact, and complicated appearance variations.

As a solution, emerging systems [8, 17, 23] leverage volumetric rendering [21], while others [13, 24, 37] use 3D Gaussians [14] as a color and volume barrier. Human priors like [19] can be used and achieve good results even on monocular setups [8, 13, 24]. Despite the impressive results, none of the mentioned methods is capable of reconstructing multiple persons in a scene.

Approaches presented in [25, 31, 35] achieve good results in performance capture and free view point video generation of humans in close interactions, but either only work with dense camera setups [31] or are limited to the setups with relatively distant performers [25] Sec. 4 or demand precise instance segmentation [35].

Concurrent work MultiPly [11] presents an approach targeting single-camera multi-human reconstruction, modeling geometry in canonical space, utilizing additional supervision signal from segmentation, which is obtained by repeatedly prompting SAM [15]. It demonstrates impressive results in model reconstruction, however novel-view synthesis is inherently limited by the single training camera and seen sides of the subject. We illustrate this limitation in the supplementary.

In our paper, we introduce a novel framework designed to capture the performance of multiple individuals engaged in close interactions. This framework excels in environments with sparse camera setups and eliminating the need for segmentation. Remarkably, it can accurately capture fine geometric details, even in scenarios where individuals are in direct physical contact, relying solely on images and roughly estimated SMPL models.

We solve the tasks of human separation and performance reconstruction directly in 3D. To achieve this, we use layered scene representation where each layer is represented by geometry and appearance neural fields. The key challenge is to separate the geometry of humans in direct physical contact without relying on segmentation. To address this problem, the following concepts were used as the basis of our approach:

i) We define a layered representation where each human has a single temporally consistent representation of shape and texture in canonical space and leverage the inverse mapping of a parametric body model to learn from deformed observations.
ii) A novel composite volume rendering technique allowing to perform rendering of multiple dense-color fields along the ray after importance sampling applied separately for each human.
iii) A quasi-background camera bounded network that allows work without external segmentation modules.
iv) Specific objective functions and SMPL-based sampling allow for faster convergence and clearer separation of the layers.

More specifically, we leverage a surface-guided approach to attain densities via the conversion method proposed in [30]. Similarly to [8], [11], we warp all sampled points into canonical space and update the human shape field dynamically. To do this, we first perform importance sampling individually for each person, similar to [30]. Then, we merge and sort the obtained points. To obtain sharp

boundaries between closely interacting humans, we individually sample density and colors from each human using the merged points obtained after importance sampling among all persons. This approach intentionally introduces ambiguities. Contrary to previous methods [17, 33], we do not predetermine which neural representation of a human is tasked with defining color and density; instead, we delegate this task to the optimization process. We also apply a novel composite rendering technique where the colors from different humans are combined in a specific way, described in 3.2. We penalize interpenetration between the different humans. Finally, we densely sample area near the humans using pre-estimated SMPL from of the shelf methods.

We show that our approach leads to clean decomposition and high-quality 3D reconstructions of human subjects even in direct physical contact. In detailed ablations we shed light on the key components of our method. Furthermore, we compare our method to existing methods operating in similar setups in geometry reconstruction tasks and in novel view synthesis task. We show on par performance with state of the art human segmentation framework and prove that our method is able to improve initially roughly estimated SMPL parameters.

To summarize our contributions:

- Approach for representing scenes with humans in close interactions
- Sampling technique and composite volume rendering.
- Combination of objectives for precise geometry reconstruction in direct physical contact without external segmentation.

2 Related work

2.1 Articulated body models

Articulated body models like [1, 19, 22] are widely used for human modeling in computer vision and computer graphics. Because of their low-dimensional parameter space and fixed topology of the underlying 3D mesh, they are well suited for learning tasks like fitting to RGB, RGBD or sparse point clouds images [2, 5, 16, 18, 20, 32]. Recent pose estimators [26] can perform really well in complicated environments with multiple humans in close interactions. Main disadvantage of this models is that they do not allow to capture clothing and sophisticated deformations.

2.2 Neural representation based methods

Seminal work [21] represents scenes with implicit fields of density and color, which are well-suited for the differentiable rendering and achieve photo-realistic view synthesis result. Another work [30] introduce SDF based volumetric rendering and novel importance sampling technique.

Neuralbody [23] anchors a set of latent codes to the vertices of the SMPL model, a deformable human body model. These latent codes are designed to capture local geometry and appearance information of the human body. For any

given frame, the method transforms the locations of the latent codes based on the human pose and uses a neural network to regress the density and color for any 3D point in space. This method enables effective integration of observations across video frames, addressing the challenge of learning from sparse views.

Faster version [7] of neuralbody significantly accelerates the optimization process for creating neural volumetric representations of dynamic humans, achieving a 100x speedup by efficient distribution of the network's representational power across different human body parts and models the 3D human deformation in a 2D domain by projecting near-surface points to neighboring regions on a parametric human model (*e.g.* SMPL).

Vid2avatar [8] parameterizes the 3D geometry and texture of the human as a pose-conditioned implicit signed-distance field and texture field in canonical pose. SMPL is used as a transport between points in canonical space and deformed space. Good results are demonstrated on monocular in the wild video scenes.

SNARF [3] represents an object by its shape and skinning weights in a canonical space. It uses a neural network to predict the occupancy probability for any 3D point in this space, incorporating pose information to capture pose-dependent local deformations. The core innovation of SNARF is its method for finding the canonical correspondences of deformed points in space. This is achieved through an iterative root-finding algorithm that solves for the points in canonical space that correspond to a given deformed point. This method deeds an adaptation to work in images setup, such us MVS preprocessing and pose estimation.

Regression-based methods that directly regress 3D surfaces from images have demonstrated compelling results [9, 29, 36]. However, they require high-quality 3D data for supervision and cannot maintain the space-time coherence of the reconstruction over the whole sequence [8].

2.3 Multi-person reconstruction

Some works dedicated specifically to multi-person reconstruction reconstruction of people in close interactions.

StNERF [33] use MVS and SiamMask tracker [10] to get instance segmentation and to assign and track bounding boxes for humans in the scene. Except color and density each layer has deformation network for mapping between canonical and current scene. The core of their volumetric rendering technique is an object-aware volume rendering scheme which involves rendering each dynamic entity (or layer) separately and then compositing them together based on their spatial relationships.

Other work [25] represents each person as a separate layer similar to StNERF. But also utilise the [23] approach and register features in SMPL nodes. They layered rendering uniformly sample from axis-aligned bounding boxes of each person, then query layers separately and finally apply standard formulation for rendering [21]. Method shows good results but work in relatively distant persons.

Deep multicapture approach presented in [35] applies the direct regression for multi-person multi-camera setup. An attention-aware module is designed

to obtain the fine-grained geometry from mutli-view images. Additionally, the paper proposes a temporal fusion method to enhance the consistency of moving character reconstructions across video frames. We will show that we outperform this method in case of direct physical contact and close interactions even without segmentation and on a roughly estimated SMPL.

HI4D [31] introduces a novel approach and dataset for analyzing close human-human interactions with prolonged contact. Firstly, they fit snarf to represent each person separately. Secondly, they fit combined persons to point cloud obtained from dense camera setup. Then method employs an iterative process that alternates between optimizing pose and refining surface details. Method performs well but requires dense camera setup.

3 Method

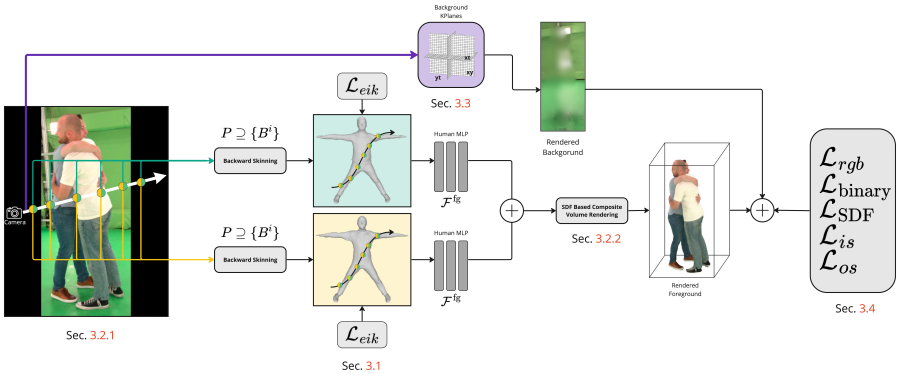


Fig. 1: We perform importance sampling separately for each human along the ray for the inner volume and merge the sampled points with distance sorting. After that, we query the geometry and color networks of each person with these points. When density and colors are queried, we use our composite volumetric rendering approach to get the foreground color and combine with learned background.

We present MV2MP, a novel framework for detailed geometry and appearance reconstruction of multiple people from sparse camera setups. The overview of our method is schematically illustrated in Figure 1. Reconstructing multiple people from a short video without prior geometry knowledge is challenging due to complex human movement and significant occlusions. To address these challenges, we first establish a unified, temporally consistent layered representation of humans 3.1. This layered neural representation is learned from images through our tailored composite volume rendering technique 3.2. Thirdly, we add quasi-background layers for each camera, eliminating the need for any external tools

for segmentation 3.3. Finally, specific objective functions 3.4 and SMPL-based sampling 3.2 allow us to achieve spatially coherent high-quality 3D reconstructions of people, as well as enhancements in SMPL models and segmentation.

3.1 Layered person representation

We represent each person $p = 1, \ldots, P$, with specific layer which may interleave. Similar to [8] we use canonical and deformed spaces for each person. To obtain color and density values in deformed space we have to know those values in canonical space as well as mapping from canonical to deformed space.

Canonical human representation Each individual in the scene is represented by an implicit signed-distance field (SDF) for the 3D shape and a texture field for the appearance, both defined in the canonical space. Specifically, we describe the geometry and appearance of each person p in canonical space using a neural network f^p. This network predicts the signed distance s^p and the color c^p at the query point x_c as follows:

$$c^p, s^p = f(x_c^p, \theta^p), \tag{1}$$

where θ^p represents the person pose parameters, which are concatenated with x_c^p to capture pose-dependent surface deformations. For ease of notation, we use $f_c^p(\cdot)$ and $f_s^p(\cdot)$ to refer separately to the network outputs c^p and s^p, .

Mapping to deformed space During volumetric rendering we are interested in obtaining color and signed distance in deformed space, i.e. person specific pose at the specific time. SMPL-based Linear Blend Skinning (LBS) is used to obtain the mapping from canonical space to deformed space. Consider the transformation matrix B_i associated with joint j_i within the set $(1, \ldots, n_b)$, which is formulated based on the pose θ of the body. Here n_b is the total count of bone elements involved in the deformation process. We map a point $\mathbf{x}^c$ in the canonical form to a transformed point $\mathbf{x}^d$ as a linear combination:

$$\mathbf{x}^d = \sum_{i=1}^{n_b} w_i B_i \mathbf{x}^c. \tag{2}$$

To obtain the original canonical position $\mathbf{x}^c$ from a transformed point $\mathbf{x}^d$, we employ the inverse operation of the equation above:

$$\mathbf{x}^c = \left(\sum_{i=1}^{n_b} w_i B_i \right)^{-1} \mathbf{x}^d. \tag{3}$$

The function $w(\cdot) = w_1, \ldots, w_{n_b}$ is indicative of the weight distribution across the skinning process for $\mathbf{x}(\cdot)$. It should be noted that the transformed locations $\mathbf{x}_d$ are correlated with the posture θ of the body. This relationship is determined by averaging the skinning weights of the closest SMPL model vertices like in [8, 28] which are in turn modulated by the distances between points in deformed space.

3.2 Compositional volume rendering

Our compositional volume rendering differs from standard one in several aspects described below.

Ray Sampling Our architecture does not require full background reconstruction; instead, we reconstruct only the quasi-background and benefit from the contrast between the foreground and background. We project the pre-estimated SMPL model onto the canvas and dilate it. Consequently, we cast rays only in the areas of interest, specifically targeting people. By that we do not spend rays to train background, which is especially useful when dealing with scenes where humans are far from each other.

Point Sampling Firstly we perform importance sampling separately for each human p for each ray r following [30] and obtain $x^p_{d,1}, \ldots, x^p_{d,N}$, where N is the number of the sampled points in each ray r. Then we merge these points and sort according to the distance from camera and obtain set of points $M = x_{m,1} \ldots x_{m,N*P}$. These points serve as hard samples preventing the density leakage from the other persons and used as a domain for the interpenetration loss described further.

Rendering Each point from M is converted to the person p canonical space, followed by the extraction of colors c^p_i and densities σ^p_i:

$$\sigma^p_i = \sigma\left(f^p_s\left(T^{-1}_{\text{SMPL}}\left(x_{m,i}, \theta^p\right), \theta^p\right)\right) \tag{4}$$

$$c^p_i = f^p_c\left(T^{-1}_{\text{SMPL}}\left(x_{m,i}, \theta^p\right), \theta^p, \ldots\right) \tag{5}$$

where $\sigma(\cdot)$ is the scaled Laplace distribution's Cumulative Distribution Function defined in [30] and T^{-1}_{SMPL} is defined in Equation 3.

Standard volume is not feasible rendering because there are several density and color signals per point $x_{m,1}$ and T_{SMPL}. We merge colors and densities from different human bodies and obtain a final persons color on the ray r using following batch of formulations:

$$C_r = \sum_{p=1}^{P} \sum_{i=1}^{N*P} w^p_i \cdot c^p_i + bg_r \tag{6}$$

$$w^p_i = \left(1 - \exp\left(\Delta x_i \sigma^p_i\right)\right) \cdot \exp\left(-\sum_{j=1}^{i} \Delta x_i \cdot \sum_{k=1}^{P} \sigma^k_i\right) \tag{7}$$

where, Δx_i is the length of the i-th segment between two adjacent sampled points in M and bg_r is the color of background described in Section 3.3.

The main idea behind equations is that we sum the densities of different individuals for calculating transparency in front of the point $x_{m,i}$. We calculate the alphas for the given human using only the density of segment of specific human ignoring others.

3.3 Background Description and Structure

In seminal work [8] the background is represented using method similar to NeRF++ [34]. This approach is resource demanding and may struggle with modeling of the dynamic background.

Instead, we use image plane coordinates (u_i, v_i) of the casted ray r and camera index to query a background color $C_{r,kp}$ from designated network based on K-planes [6]. The final background color bg_r is obtained multiplying by the remaining alpha value $bg_r = \alpha_r \cdot C_{r,kp}$, where $\alpha_r = \left(1 - \sum_{p=1}^{P} \sum_{i=1}^{N*P} w_i^p\right)$.

This method ensures that during training, it is not beneficial for k-planes to learn foreground colors, as the person movements would degrade the reconstruction of background. Similarly, it is not advantageous for the human networks to learn the background, as this is penalized by the adjacent frames and camera views.

3.4 Loss Functions

The most common objective for neural rendering models is the pixel-wise photometric loss: L_{rgb} [8] that forces renderings to reconstruct input images. For SDF-based rendering, a regularizer L_{eikonal} [27] is usually used to constrain the implicit geometry to satisfy the Eikonal equation: $||\nabla \text{sdf}||_2 = 1$.

Additionally, we apply binary cross-entropy loss on obtained per-pixel opacities, this forces opacity to be either zero or one:

$$L_{\text{BCE}} = -\left(\alpha_r \cdot \log(\alpha_r) + (1 - \alpha_r) \cdot \log(1 - \alpha_r)\right) \tag{8}$$

In the context of multi-person setups, we introduce two additional losses alongside the BCE loss: the global opacity sparseness regularization and the in-shape loss following the [8]. These loss functions are designed to refine the volumetric representation by imposing constraints on the opacity values corresponding to regions both outside and inside the human shapes.

The precise loss formulation and scheduling can be obtained in provided source code.

Interpenetration Penalty We consider the scenario where P individuals are represented by their signed distance functions $s_i^p = f_s^p\left(T_{\text{smpl}}^{-1}\left(x_{m,i}, \theta^p\right), \theta^p\right)$. We formulate a loss to minimize intersection between modeled bodies. For each pair of individuals we find points where both SDF values are negative, indicating an overlap:

$$L_{\text{SDF}} = \frac{1}{\binom{N}{2}} \sum_{p=1}^{P} \sum_{k=p+1}^{P} \sum_{i=1}^{N*P} \mathbf{1}_{\{s_i^k, s_i^p | s_i^k < 0, s_i^p < 0\}}(s_i^k, s_i^p) s_i^k \cdot s_i^p \tag{9}$$

where $\binom{N}{2}$ represents the total number of unique pairs in a scene with N individuals. This loss effectively minimizes the physical overlap between any two individuals, encouraging a more realistic and physically plausible rendering of multiple human models by penalizing intersections in their volumetric representations.

4 Experiments

We conducted our experiments on two widely-used datasets: the HI4D [31] and the CMU Panoptic [12]. The presented results provide empirical evidence that the proposed method outperforms concurrent work and consistently operates without relying on ground truth segmentation or high-quality SMPL pose estimations. Additional results for instance segmentation quality and SMPL pose refinement can be found in the in supplementary material.

4.1 Datasets and metrics

The HI4D dataset [31] provides high-resolution sequences of human interactions in various scenarios. Each timestamp within a sequence contains ground-truth textured meshes, SMPL parameters, 8 RGB images, and instance segmentation masks. We selected scenes involving close physical interactions to test the robustness of our approach. HI4D ground-truth SMPL data and instance masks are of high quality, because they were obtained using large number of cameras and depth sensors. We decided that using only these may potentially lead to overly optimistic results. Therefore, to ensure realistic evaluation, we also estimate SMPL parameters using the widely-used open-source tool [25] and generate instance segmentation masks using nearly state of the art the Mask2Former [4] model. We fit models and report metrics using both ground truth HI4D and estimated input data.

The CMU Panoptic dataset [12] contains sequences captured from multiple synchronized cameras. Although the provided sequences do not include close physical interactions, they feature scenes with multiple individuals, which allows us to assess our method's performance in multi-person scenarios. The CMU Panoptic dataset does not provide ground truth data for SMPL parameters and instance masks, so we use the same estimation tools as for the HI4D dataset. We use CMU-P dataset for qualitative comparison.

To ensure comprehensive comparison, we train models on setups with 3, 5, or 7 cameras and use from 1 to 3 cameras for testing. The train/test splits are listed in the supplementary material.

We evaluate our method using Chamfer Distance (CD) and Peak Signal-to-Noise Ratio (PSNR). Metric values are averaged across all validation cameras.

Table 1: Comparative Analysis of Mesh Reconstruction Quality. For rows marked with GT we use groundtruth SMPL and masks data from the dataset for model fitting and estimated data otherwise.

Dataset, scene	**# views**	**DeepMultiCap**	**MultiNB**		**Ours**	
		CD	CD	PSNR	CD	PSNR
HI4D, hug21, GT	7	1.76	2.25	22.02	-	-
HI4D, hug21	7	2.60	2.66	20.98	**1.48**	**23.97**
HI4D, hug21	5	2.72	2.95	20.35	**1.41**	**23.64**
HI4D, hug21	3	2.86	-	-	1.37	22.59
HI4D, yoga00, GT	7	1.95	2.08	20.43	-	-
HI4D, yoga00	7	2.94	2.07	**19.29**	**1.48**	18.13
HI4D, yoga00	5	3.11	2.46	**17.95**	**1.51**	18.35
HI4D, yoga00	3	3.39	-	-	1.63	17.56
HI4D, sidehug32	7	2.85	2.54	20.66	**1.79**	**22.89**
HI4D, sidehug32	5	2.84	2.81	19.28	**1.93**	**21.68**
HI4D, sidehug32	3	3.07	-	-	2.27	19.77

4.2 Comparisons on mesh reconstruction and novel view synthesis

We make a comparison between the proposed method and two established methods which use similar sparse multi-camera setup:

1. Method [25] which we refer as MultiNB is based on a layered neural scene representation. Originally designed for novel view synthesis and instance mask generation in multi person scenario, it allows to extract an individual mesh from the underlying neural field.
2. DeepMultiCap [35] employs a space attention-aware network to capture fine-grained body details and reconstructs each individual independently. DeepMultiCap was specifically designed for close interactions and has exhibited generalization to unseen scenes.

For our evaluations, we used the implementations as provided by the original authors. For brevity, we present the averaged numbers here; additional details can be found in the supplementary section.

The process of training for our method takes approximately 15 hours to converge on a sequence consisting of 100 frames and 8 cameras, when executed on a single A100 GPU.

Results in mesh reconstruction. Referring to Table 1, we report the performance of MultiNB, DeepMultiCap, and our method on three scenes from the HI4D dataset, featuring diverse interactions. All results were obtained using the same sets of cameras and frames. Both MultiNB and our model were trained on sequences of 100 frames. The DeepMultiCap model was used in its provided

pretrained form, with only the necessary data preparation conducted as outlined in the official instructions.

When applied to the estimated input data on the given camera setup, the performance of both MultiNB and DeepMultiCap degrades, while our method's performance remains nearly the same and outperforms competitive methods even without instance masks and high-quality SMPL models. We attribute this stability to our SMPL mesh optimization and learned background.

Figure 2 shows qualitative comparison and more examples are available in the supplementary material.

(a) DMC **(b)** MultiNB **(c)** Ours **(d)** GT

Fig. 2: Comparison of Mesh Outputs for Different Methods.

(a) MultiNB **(b)** Ours **(c)** GT

Fig. 3: Visual comparison of novel views produced by MultiNB and our methods on HI4D datset.

Results in novel view synthesis. In the novel view synthesis experiment, we compared the performance of our method with the MultiNB model. We used the average PSNR over validation cameras and the entire sequence as a metric, specifically reporting PSNR for areas covered by people. The results, detailed in Table 1, show that our models have comparable capabilities in novel view synthesis. The lower PSNR observed on the yoga00 scene may be due to the striped pattern on the actor's T-shirt, which could result from undertraining. Qualitative comparisons are shown in Figure 3 and in Figure 4.

(a) MultiNB **(b)** Ours **(c)** GT

Fig. 4: Visual comparison of novel views produced by MultiNB and our method in multi person setup on CMU-P dataset.

5 Ablation Study

In order to check contribution of different details of our method we check different aspects of performance using restricted versions of our approach:

1. we check how our method performs if we do not model quasi-background, i.e. we use segmentation masks from off-the-shelf segmentator and train the model to reconstruct only the foreground, we substitute mesh-based labels in In-Shape loss and in Opacity Sparse loss with ones provided by off-the-shelf segmentation mask.
2. as we find, that our method can tune objects' shapes and poses, we check if this has any importance in terms of final metrics.
3. one of our novelties is compositional rendering: we first find query points from objects on the scene, and query each object's SDF to find each object's density contributions to integrate them all into final color. Here we switch off querying foreign SDFs on points obtained for object i. i.e. we query only $i'th$ SDFs for those points.

We perform our ablation on HI4D since it provides accurate meshes, segmentation mask, humans skeletons so we can compute PSNRs, IOUs, Chamfer Distances, Skeleton Per Joint Distances.

Contribution of qausi-background modeling.

We took our 7 train cameras setups from HI4D scene and stripped the background modeling capability, in Table 2 one can see, that we get comparable CD and IOU metrics, while improving on masked PSNR by around 1 point. See also Fig. 5 for visual comparison.

Contribution of tuning of shapes and poses.

For ablation, we switch off SMPL tuning as well, to check if our model benefits from more accurate coarse geometry, provided by body model. We check Chamfer Distance w.r.t. ground-truth meshes, IOU with relation w.r.t semantic masks, MPJPE w.r.t GT joints, PSNR. According to Table 3 we benefit hugely from introducing SMPL shapes and poses finetuning. We hypothesize, that this has to do with the fact, that our model relates on skinning abilities of SMPL geometry, therefore providing more accurate coarse geometry leads to more fi-

Table 2: Metric comparison for quasi background modeling ablation

dataset, method ↓	**CD**	**IOU**	**PSNR**
HI4D: yoga00, ours	1.480	0.934	18.138
HI4D: yoga00, ours w/o bg	1.462	0.940	17.225
HI4D: hug21, ours	1.483	0.937	23.974
HI4D: hug21, ours w/o bg	1.334	0.952	23.072
HI4D: sidehug32, ours	1.797	0.928	22.885
HI4D: sidehug32, ours w/o bg	1.704	0.940	21.722

Table 3: Metric comparison for body model fitting ablation. Our method benefits hugely from SMPL fine-tuning, PSNR gains are in order of 1 point, Chamfer Distance decreased by around 25% on each scene

dataset, method ↓	**CD**	**IOU**	**MPJPE**	**PSNR**
HI4D: yoga00, ours	1.480	0.934	0.073	18.138
HI4D: yoga00, ours w/o SMPL Tuning	1.832	0.907	0.073	17.184
HI4D: hug21, ours	1.483	0.937	0.040	23.974
HI4D: hug21, ours w/o SMPL Tuning	2.079	0.898	0.045	21.768
HI4D: sidehug32, ours	1.797	0.928	0.050	22.885
HI4D: sidehug32, ours w/o SMPL Tuning	2.201	0.910	0.055	20.986

nal consistent geometry, resulting in less canonical geometry bumps and giving better reconstruction results.

Compositional rendering contribution.

By introducing compositional rendering, i.e. using foreign points along the ray to query each SDF to accumulate all densities, we observe slight increase in masked PSNR, while maintaining both CD, MPJPE, IOU comparable to method without compositional rendering. We as well trade off training speed, as we need to query n times more SDF network passes (n - number of persons on the scene). See Table 4 for details.

6 Conclusion

In this paper, we presented the approach suitable for novel view synthesis and reconstruction of intricate multi-human interactions which relies on a sparse set of calibrated cameras. Utilization of the canonical SDF enables us to effectively extract the mesh of an individual under occlusions, provided they were visible in some frames. We have conducted evaluation and show comparative perfor-

Table 4: Metric comparison for compositional rendering ablation

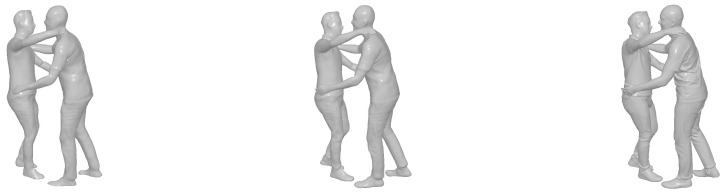
dataset, method ↓	**PSNR**
HI4D: yoga00, ours	18.138
HI4D: yoga00, ours w/o compositonal rendering	17.566
HI4D: hug21, ours	23.974
HI4D: hug21, ours w/o compositonal rendering	23.728
HI4D: sidehug32, ours	22.885
HI4D: sidehug32, ours w/o compositonal rendering	22.502

(a) Full **(b)** w/o background **(c)** Ground truth

Fig. 5: Visual comparison of novel views produced by base and no-quasi-background methods. Notice slightly less detailed hands while using ablated method.

mance in terms of PSNR and outperform existing methods in terms of mesh reconstruction accuracy.

Limitations and future work. At first, our method struggle with fine details, e.g. nuanced elements such as individual fingers. The integration of advanced parametric body models like SMPLX could potentially address this shortcomings. Secondly, the reliance on SMPL models can be restrictive. Lastly, the current scope of our method is limited to reconstructing human figures.



(a) Without tuning **(b)** With SMPL tuning **(c)** Ground truth

Fig. 6: Visual comparison of mesh renders with and without SMPL tuning. Notice how hands are being put more accurately when using SMPL tuning.

References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers, pp. 408–416 (2005)
2. Bogo, F., Black, M.J., Loper, M., Romero, J.: Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In: Proceedings of the IEEE international conference on computer vision. pp. 2300–2308 (2015)
3. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11594–11604 (2021)
4. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation (2022)
5. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. pp. 20–40. Springer (2020)
6. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479–12488 (2023)
7. Geng, C., Peng, S., Xu, Z., Bao, H., Zhou, X.: Learning neural volumetric representations of dynamic humans in minutes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8759–8770 (2023)
8. 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)
9. He, T., Collomosse, J., Jin, H., Soatto, S.: Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. Advances in Neural Information Processing Systems **33**, 9276–9287 (2020)
10. Hu, W., Wang, Q., Zhang, L., Bertinetto, L., Torr, P.H.: Siammask: A framework for fast online object tracking and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3072–3089 (2023)
11. Jiang, Z., et al.: Multiply: Reconstruction of multiple people from monocular video in the wild. In: Proceedings of the IEEE/CVF CVPR (June 2024)
12. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T.S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social interaction capture. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
13. Jung, H., Brasch, N., Song, J., Perez-Pellitero, E., Zhou, Y., Li, Z., Navab, N., Busam, B.: Deformable 3d gaussian splatting for animatable human avatars. arXiv preprint arXiv:2312.15059 (2023)
14. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)

16. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5253–5263 (2020)
17. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. Advances in Neural Information Processing Systems **34**, 24741–24752 (2021)
18. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3383–3393 (2021)
19. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
20. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
21. Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J., Ramamoorthi, R., Nerf, R.N.: Representing scenes as neural radiance fields for view synthesis., 2021, 65. DOI: https://doi. org/10.1145/3503250 pp. 99–106
22. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)
23. Peng, S., Geng, C., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Zhou, X., Bao, H.: Implicit neural representations with structured latent codes for human body modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
24. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. arXiv preprint arXiv:2312.09228 (2023)
25. Shuai, Q., Geng, C., Fang, Q., Peng, S., Shen, W., Zhou, X., Bao, H.: Novel view synthesis of human interactions from sparse multi-view videos. In: ACM SIGGRAPH 2022 Conference Proceedings. SIGGRAPH '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3528233.3530704, https://doi.org/10.1145/3528233.3530704
26. Shuai, Q., Yu, Z., Zhou, Z., Fan, L., Yang, H., Yang, C., Zhou, X.: Reconstructing close human interactions from multiple views. ACM Trans. Graph. **42**(6) (dec 2023). https://doi.org/10.1145/3618336, https://doi.org/10.1145/3618336
27. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
28. Xiang, T., Sun, A., Delp, S., Kozuka, K., Fei-Fei, L., Adeli, E.: Wild2avatar: Rendering humans behind occlusions. arXiv preprint arXiv:2401.00431 (2023)
29. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: Icon: Implicit clothed humans obtained from normals. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13286–13296. IEEE (2022)
30. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems **34**, 4805–4815 (2021)
31. Yin, Y., Guo, C., Kaufmann, M., Zarate, J.J., Song, J., Hilliges, O.: Hi4d: 4d instance segmentation of close human interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17016–17027 (2023)

32. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7287–7296 (2018)
33. Zhang, J., Liu, X., Ye, X., Zhao, F., Zhang, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Editable free-viewpoint video using a layered neural representation. ACM Transactions on Graphics (TOG) **40**(4), 1–18 (2021)
34. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv preprint arXiv:2010.07492 (2020)
35. Zheng, Y., Shao, R., Zhang, Y., Yu, T., Zheng, Z., Dai, Q., Liu, Y.: Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6239–6249 (2021)
36. 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 pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3050505
37. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3d gaussian avatars. arXiv preprint arXiv:2311.08581 (2023)