Hi4D: 4D Instance Segmentation of Close Human Interaction

Yifei Yin Chen Guo Manuel Kaufmann Juan Jose Zarate Jie Song Otmar Hilliges
ETH Zurich

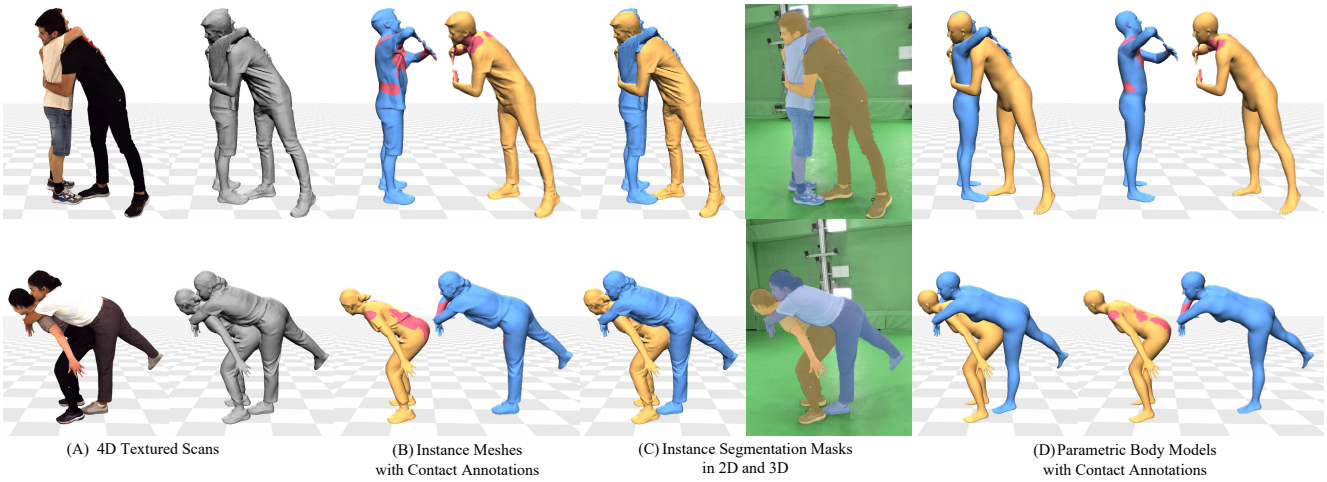

Figure 1. **Method and dataset.** We propose a method that leverages personalized human avatars to segment merged (A) 4D textured scans of multiple closely interacting humans. Based on the (B) instance meshes (with vertex-level contact annotations) we obtained in 3D space, we further provide (C) instance segmentation masks in 2D and 3D, (D) registered parametric body models with contact annotations.

Abstract

We propose Hi4D, a method and dataset for the automatic analysis of physically close human-human interaction under prolonged contact. Robustly disentangling several in-contact subjects is a challenging task due to occlusions and complex shapes. Hence, existing multi-view systems typically fuse 3D surfaces of close subjects into a single, connected mesh. To address this issue we leverage i) individually fitted neural implicit avatars; ii) an alternating optimization scheme that refines pose and surface through periods of close proximity; and iii) thus segment the fused raw scans into individual instances. From these instances we compile Hi4D dataset of 4D textured scans of 20 subject pairs, 100 sequences, and a total of more than 11K frames. Hi4D contains rich interaction-centric annotations in 2D and 3D alongside accurately registered parametric body models. We define varied human pose and shape estimation tasks on this dataset and provide results from state-of-the-art methods on these benchmarks. Hi4D dataset can be found at https://ait.ethz.ch/Hi4D.

1. Introduction

While computer vision systems have made rapid progress in estimating the 3D body pose and shape of individuals and well-spaced groups, currently there are no methods that can robustly disentangle and reconstruct *closely* interacting people. This is in part due to the lack of suitable datasets. While some 3D datasets exist that contain human-human interactions, like ExPI [72] and CHI3D [22], they typically lack high-fidelity dynamic textured geometry, do not always provide registered parametric body models and do not always provide rich contact information and are therefore not well suited to study closely interacting people.

Taking a first step towards future AI systems that are able to interpret the interactions of multiple humans in close physical interaction and under strong occlusion, we propose a method and dataset that enables the study of this new setting. Specifically, we propose Hi4D, a comprehensive dataset that contains segmented, yet complete 4D textured geometry of *closely* interacting humans, alongside corresponding registered parametric human models, instance segmentation masks in 2D and 3D, and vertex-level contact annotations (see Fig. 1). To enable research to-

Dataset	Multi-view Images	Temporal	Reference Data Modalities			
			3D Pose Format	Textured Scans	Contact Annotations	Instance Masks
ShakeFive2 [23]		✓	Joint Positions			
MuPoTS-3D [53]	✓	✓	Joint Positions			
ExPI [72]	✓	✓	Joint Positions	✓		
MultiHuman [77]			Parametric Body Model†	✓		
CHI3D [22]	✓	✓	Parametric Body Model		region-level (631 events)	
Hi4D (Ours)	✓	✓	Parametric Body Model	✓	vertex-level (> 6K events)	✓

Table 1. **Comparison of datasets containing close human interaction.** †In [77] registrations are not considered as ground-truth.

wards automated analysis of close human interactions, we contribute experimental protocols for computer vision tasks that are enabled by Hi4D.

Capturing such a dataset and the corresponding annotations is a very challenging endeavor in itself. While multi-view, volumetric capture setups can reconstruct high-quality 4D textured geometry of individual subjects, even modern multi-view systems typically fuse 3D surfaces of spatially proximal subjects into a single, connected mesh (see Fig. 1, A). Thus deriving and maintaining complete, per subject 4D surface geometry, parametric body registration, and contact information from such reconstructions is non-trivial. In contrast to the case of rigid objects, simple tracking schemes fail due to very complex articulations and thus strong changes in terms of geometry. Moreover, contact itself will further deform the shape.

To address these problems, we propose a novel method to track and segment the 4D surface of multiple closely interacting people through extended periods of dynamic physical contact. Our key idea is to make use of emerging neural implicit surface representations for articulated shapes, specifically SNARF [13], and create personalized human avatars of each individual (see Fig. 2, A). These avatars then serve as strong personalized priors to track and thus segment the fused geometry of multiple interacting people (see Fig. 2, B). To this end, we alternate between pose optimization and shape refinement (see Fig. 3). The optimized pose and refined surfaces yield precise segmentations of the merged input geometry. The tracked 3D instances (Fig. 1, B) then provide 2D and 3D instance masks (Fig. 1, C), vertex-level contact annotations (Fig. 1, B), and can be used to register parametric human models (Fig. 1, D).

Equipped with this method, we capture Hi4D, which stands for Humans interacting in 4D, a dataset of humans in close physical interaction alongside high-quality 4D annotations. The dataset contains 20 pairs of subjects (24 male, 16 female), and 100 sequences with more than 11K frames. To our best knowledge, ours is the first dataset containing rich interaction-centric annotations and high-quality 4D textured geometry of closely interacting humans.

To provide baselines for future work, we evaluate several state-of-the-art methods for multi-person pose and shape modeling from images on Hi4D in different settings such as monocular and multi-view human pose estimation and detailed geometry reconstruction. Our baseline experiments show that our dataset provides diverse and challenging benchmarks, opening up new directions for research. In summary, we contribute:

- A novel method based on implicit avatars to track and segment 4D scans of closely interacting humans.
- Hi4D, a dataset of 4D textured scans with corresponding multi-view RGB images, parametric body models, instance segmentation masks and vertex-level contact.
- Several experimental protocols for computer vision tasks in the close human interaction setting.

2. Related Work

Instance Segmentation. Most works that tackle human instance segmentation [45–47, 63, 68] or object detection in general [10, 24, 28, 39, 40, 48, 58, 59] are only applicable to the 2D domain. These methods do not naïvely transfer to the 3D domain [29]. For 3D instance segmentation, previous work predominantly focuses on scene understanding that does not include humans [16, 20]. Thus, 3D instance segmentation of humans in close interactions is a relatively under-explored task. Unlike static objects, humans undergo articulated motion, interact dynamically with surroundings, and cannot be represented by simple geometric primitives, which makes human-centric 3D instance segmentation inherently challenging. To address this challenging problem, we create personalized human avatars of each individual and subsequently leverage them as priors to track and segment closely interacting 3D humans.

Human Pose and Shape Modeling. Explicit body models [3, 37, 50, 57, 74] 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 images [15, 38, 41, 43, 44, 57, 66], RGB-D [8, 14, 75], or sparse point clouds [49, 52]. Yet, the fixed 3D topology limits the maximum resolution and the expressive power to represent individual features and clothing. While there have been efforts to alleviate this [1, 2, 6, 25, 51], recent attention has turned to the use of implicit representations to tackle these limitations. [13, 17, 62, 70] have shown promising results for modeling articulated clothed human

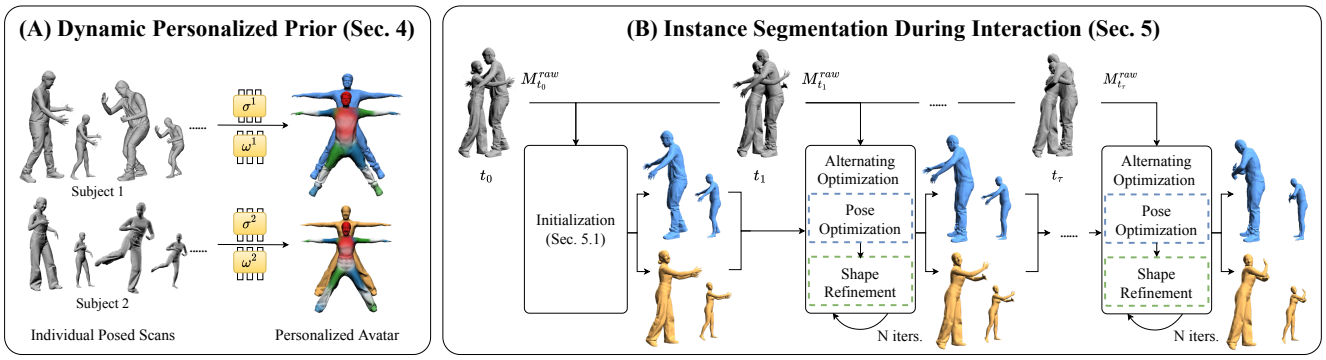

Figure 2. **Method overview.** (A) Dynamic Personalized Priors: We build individual personalized implicit avatars from 4D posed scans of each subject by modeling shape and deformation fields in canonical space following [13]. (B) Instance Segmentation During Interaction: We then leverage the pre-built individual avatars to track and segment the raw 4D scans of multiple closely interacting people through extended periods of dynamic physical contact by optimizing pose and shape in an alternating manner (*cf*. Fig. 3 for more details).

bodies in 3D. Among these, SNARF [13] achieves state-of-the-art results and shows good generalization to unseen poses. We thus use it as a building block in our method, which - according to our ablations - is the more suitable choice than an explicit body model.

Multi-Person Pose and Shape Estimation. Compared to the remarkable progress that has been made in estimating pose and shape of single humans from images or videos [26, 32, 33, 36, 38, 43, 44, 60, 61, 65, 73, 78], not much attention has been paid to multi-person pose and shape estimation for *closely* interacting humans. Multi-person estimators, *e.g*. [18, 19, 34, 41, 42, 55, 66, 67, 71], mainly deal with the case where people are far away from each other and do not interact naturally in close range. While the works of [76, 77] show more closely interacting people, the focus lies more on occlusions caused by this scenario and the actual contact between body parts is often limited. [22] study closer human interactions in a similar setting to ours, but without textured scans. Their method to disentangle interacting people is fundamentally different from ours and heavily relies on manual annotations, which our method is able to avoid through the designed optimization schema.

Close Human Interaction Datasets. There are several contact-related datasets focusing on how humans interact with objects or static scenes [7, 21, 27, 31, 69]. None of them considers close interactions between dynamic humans. Of the datasets containing human-human interactions [22, 23, 30, 35, 53, 72, 77] the most recent ones with close human interactions are summarized in Tab. 1. ShakeFive2 [23] and MuPoTS-3D [53] only provide 3D joint locations as reference data, lacking body shape information. The most related dataset to ours is CHI3D [22], which employs a motion capture system to fit parametric human models of at most one actor at a time. CHI3D only provides contact labels at body region-level and only for 631 frames, whereas we provide vertex-level annotations at more than 6K instances. Furthermore, CHI3D does not contain textured scans, which are crucial to evaluate surface reconstruction tasks. MultiHuman [77] provides textured scans of interacting people, but only of 453 static frames and without ground-truth level body model registrations. ExPI [72] contains dynamic textured meshes in addition to 3D joint locations, but misses instance masks along with body model registrations and contact information. Moreover, there are only two pairs of dance actors in ExPI [72], thus lacking in subject and clothing diversity. In contrast, our dataset Hi4D encompasses a rich set of data modalities for closely interacting humans.

3. Approach Overview

Vision-based disentanglement of in-contact subjects is a challenging task due to strong occlusions and a priori unknown geometries. Hence, multi-view systems typically fuse 3D surfaces of close subjects into a single, connected mesh. Here we detail our method to segment 4D scans of closely interacting people to obtain instance-level annotations. Our method makes use of three components: i) we fit individual neural implicit avatars to frames without contact (*cf*. Fig. 2, A & Sec. 4); ii) these serve as personalized priors in an alternating optimization scheme that refines pose (Sec. 5.2) and surface (Sec. 5.3) through periods of close proximity; and iii) thus segment the fused 4D raw scans into individual instances (*cf*. Fig. 2, B & Sec. 5).

4. Dynamic Personalized Prior

To build personalized priors, we first capture 4D scans for each subject during dynamic motion. Minimally clothed parametric body models (here, SMPL) are registered to these scans (Sec. 4.1). Next, detailed avatars are learned for each subject (Sec. 4.2). We leverage these to alleviate instance ambiguities during close interaction (Sec. 5).

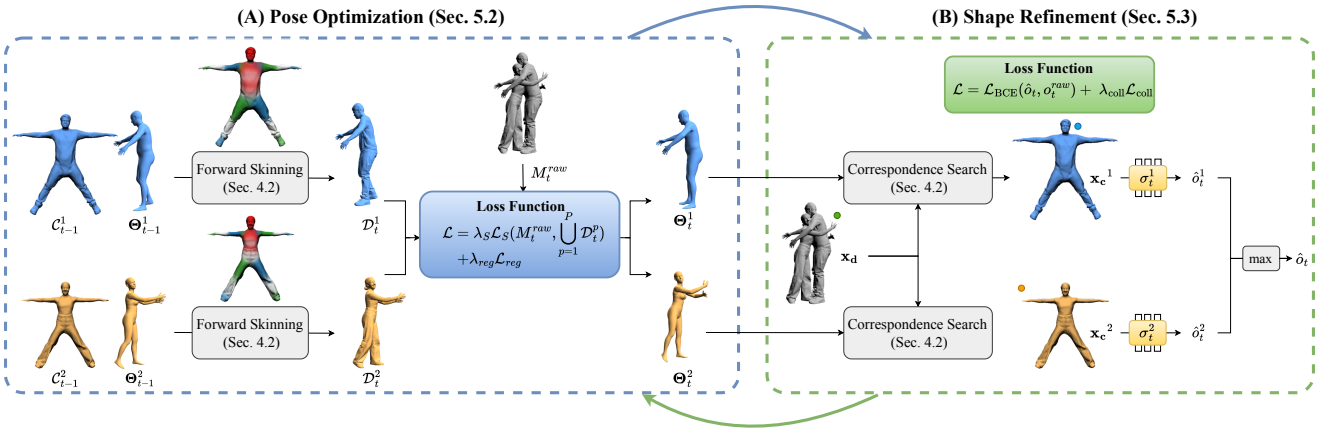

Figure 3. **Alternating optimization.** (A) Given personalized avatars for each subject (Fig. 2), we jointly optimize the poses of each subject $\mathbf{\Theta}_t^p$ using a surface energy term. (B) With $\mathbf{\Theta}_t^p$ and merged raw scans M_t^{raw}, we refine the shape network weights σ_t^p of the individual avatars on the fly to maximally preserve the details and model contact-aware deformations. We alternate between optimizing (A) and (B) for N iterations. Afterwards, the optimized $\mathbf{\Theta}_t^p$ and refined σ_t^p serve as initialization for the optimization process at the next frame $t+1$.

4.1. Parametric Body Model Fitting

We register SMPL [50] to the individual scans to represent the underlying body and its pose. SMPL is defined as a differentiable function that maps shape parameters $\beta \in \mathbb{R}^{10}$, pose parameters $\theta \in \mathbb{R}^{72}$ and translation $t \in \mathbb{R}^3$ to a body mesh $\mathcal{M}$ with 6890 vertices. Registering the SMPL model is formulated as an energy minimization problem over body shape, pose, and translation parameters:

$$E(\beta, \theta, t) = \lambda_S E_S + \lambda_J E_J + \lambda_\theta E_\theta + \lambda_\beta E_\beta \,, \quad (1)$$

where E_S denotes the bi-directional distances between the SMPL mesh and the corresponding scan, and E_J is a 3D keypoint energy between regressed SMPL 3D joints and 3D keypoints which are obtained via triangulation of 2D keypoints detected in the multi-view RGB images [11]. E_θ and E_β are prior terms to constrain human pose and shape (*cf*. [9]). Each term is weighted with a corresponding weight λ. See Sup. Mat. for more details.

To obtain high-quality registrations, the body shape β is estimated in advance from a minimally clothed static scan for each subject following [4, 5]. For registering SMPL to clothed scans, we keep β fixed in Eq. (1). For brevity, we summarize the parameters as $\mathbf{\Theta} = (\beta, \theta, t)$.

4.2. Human Avatar Learning

Neural implicit surfaces can be used to represent articulated human bodies [13, 17, 62, 70]. We leverage SNARF [13] for its generalization to unseen and challenging poses. Following [13] we use two neural fields to model shape and deformation in canonical space:

- **Shape Field:** f_σ is used to predict the occupancy probability $\hat{o}\left(\mathbf{x}_c, \mathbf{\Theta}\right)$ of any 3D point $\boldsymbol{x}_c$ in canonical space, where 1 is defined as inside and 0 as outside. The SMPL pose parameters are provided as an input to model pose-dependent deformations. The canonical shape is implicitly defined as the 0.5 level set $\mathcal{C} = \{\, \boldsymbol{x}_c \mid f_\sigma(\boldsymbol{x}_c, \mathbf{\Theta}) = 0.5 \,\}$.
- **Deformation Field:** $\mathbf{w}_\omega$ denotes a person-specific, canonical deformation field. It transforms the acquired shape to a desired pose $\mathbf{\Theta}$ via linear blend skinning (LBS) with learned deformation field.

Correspondence Search. Given a query point sampled in deformed space $\boldsymbol{x}_d$, [13] determines its correspondence $\boldsymbol{x}_c$ in canonical space via iterative root finding such that it satisfies the forward skinning function $\boldsymbol{x}_d = \boldsymbol{d}_\omega(\boldsymbol{x_c}, \mathbf{\Theta})$.
Training Losses. The implicit model is then trained via binary cross entropy $\mathcal{L}_{BCE}\left(\hat{o}\left(\mathbf{x}_d, \mathbf{\Theta}\right), o^{raw}\left(\mathbf{x}_d\right)\right)$, formed between the predicted occupancy $\hat{o}\left(\mathbf{x}_d, \mathbf{\Theta}\right)$ and the ground-truth occupancy $o^{raw}\left(\mathbf{x}_d\right)$ of points $\mathbf{x}_d$ in deformed space.
Mesh Extraction. We use *Multiresolution IsoSurface Extraction* (MISE) [54] to extract meshes $\mathcal{C}$ from the continuous occupancy fields in canonical space:

$$\mathcal{C} = MISE(f_\sigma, \mathbf{\Theta}). \quad (2)$$

The canonical shape can be deformed to posed space via linear blend skinning using the learned deformation field $\mathbf{w}_\omega$. For brevity, we denote this deformation as

$$\mathcal{D} = LBS(\mathcal{C}, \mathbf{w}_\omega, \mathbf{\Theta}). \quad (3)$$

5. Instance Segmentation During Interaction

Given pre-built individual implicit avatars obtained in Sec. 4, our goal is to track and segment the 4D scans through extended periods of dynamic physical contact. To

Algorithm 1 Alternating optimization to estimate SMPL parameters $\mathbf{\Theta}_t^p$ and refine shape network weights σ_t^p for each frame $t_1 \leqslant t \leqslant t_\tau$ in which contact happens.

$\mathbf{\Theta}_{t_0}^p \leftarrow$ Initial SMPL parameters of subject p at t_0
$\sigma_{t_0}^p \leftarrow$ Initial shape network weights of subject p at t_0
for $t = t_1, \ldots, t_\tau$ **do**
  $\mathbf{\Theta}_t^{p(0)} \leftarrow \mathbf{\Theta}_{t-1}^p$
  $\sigma_t^{p(0)} \leftarrow \sigma_{t-1}^p$
  for $n = 1, \ldots, N$ **do**
    $\mathbf{\Theta}_t^{p(n)} \leftarrow$ Pose Optimization (Eq. (4), Sec. 5.2)
    $\sigma_t^{p(n)} \leftarrow$ Shape Refinement (Eq. (7), Sec. 5.3)
  end for
  $\mathbf{\Theta}_t^p \leftarrow \mathbf{\Theta}_t^{p(N)}$
  $\sigma_t^p \leftarrow \sigma_t^{p(N)}$
end for

this end, we leverage the avatars as priors to compensate for ambiguities caused by contact. The process includes the following steps: i) given the last frame t_0 before contact, we initialize pose parameters from the still separated scans (Sec. 5.1); ii) starting from the first frame with contact, we then jointly refine the pose parameters $\mathbf{\Theta}_t^p$ for all P subjects $p \in \{1, \ldots, P\}$ in the scene via minimization of a surface energy term (Sec. 5.2); iii) we further refine the implicit shape network weights σ_t^p on the fly throughout the interaction sequence, using the optimized poses and raw scans which at this point are merged. This leads to maximal preservation of details and allows modelling of contact-aware deformations (Sec. 5.3). Steps ii) and iii) are performed in an alternating fashion for N steps. To be noted, this is a tracking process over time. The method is illustrated in Fig. 3 and Alg. 1.

5.1. Initialization

We denote the last frame without physical contact by t_0 and denote the last frame with contact by t_τ. We register the SMPL model to separated scans to obtain the initial pose parameters $\mathbf{\Theta}_0^p$ for each subject. We further use $\mathbf{\Theta}_0^p$ and the corresponding avatar to extract the canonical shape $\mathcal{C}_0^p$ for each subject p via Eq. (2). During frames $t \in \{t_1, \ldots, t_\tau\}$ with physical contact, the raw scan M_t^{raw} is fused together. To track through this period, we initialize the shape $\mathcal{C}_t^p$ for each subject from the last frame $\mathbf{\Theta}_{t-1}^p$.

5.2. Pose Optimization

To obtain the SMPL parameters $\mathbf{\Theta}_t^p$ for contact frames, we optimize the following objective:

$$\mathcal{L} = \lambda_{s2m}\mathcal{L}_{s2m}(M_t^{raw}, \bigcup_{p=1}^{P} \mathcal{D}_t^p) + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{4}$$

where $\mathcal{L}_{s2m}$ encourages the union of the posed meshes $\mathcal{D}_t^p = LBS(\mathcal{C}_t^p, \mathbf{w}_{\sigma_w}, \mathbf{\Theta}_t^p)$ (Eq. (3)) to align with the fused input scan M_t^{raw} by optimizing the SMPL parameters of each subject $\mathbf{\Theta}_t^p$ jointly. $\mathcal{L}_{\text{reg}}$ is a regularization term:

$$\mathcal{L}_{\text{reg}} = \sum_{p=1}^{P} \mathcal{L}_{s2m}(\mathcal{D}_t^p, \mathcal{M}_t^p) + \lambda_\Theta \mathcal{L}_\Theta(\mathbf{\Theta}_t^p). \tag{5}$$

The term $\mathcal{L}_{s2m}(\mathcal{D}_t^p, \mathcal{M}^p)$ ensures that the SMPL template $\mathcal{M}^p$ aligns well with each subject's deformed surface and $\mathcal{L}_\Theta$ is a prior penalizing unrealistic human poses (*cf*. [9]) and $\lambda_{(\cdot)}$ denote the corresponding weights. The scan-to-mesh loss term $\mathcal{L}_{s2m}$ is defined in Supp. Mat.

5.3. Shape Refinement

After the pose optimization stage, we refine the shape networks $f_{\sigma_t^p}$ of each avatar to retain high-frequency details and to model contact-induced deformations. To achieve this, we sample points $\boldsymbol{x}_d$ on the raw scan M_t^{raw} and find canonical correspondences $\boldsymbol{x}_c^p$ per subject, given the optimized poses $\mathbf{\Theta}_t^p$. For each $\boldsymbol{x}_c$, the shape network $f_{\sigma_t^p}$ predicts the subject-specific occupancy $\hat{o}_t^p$ (Sec. 4.2). The final occupancy prediction $\hat{o}_t$ of the sampled query point $\boldsymbol{x}_d$ is composited as the union over the individual predictions $\hat{o}_t^p$:

$$\hat{o}_t = \max_{p \in \{1,\ldots,P\}} [\hat{o}_t^p] = \max_{p \in \{1,\ldots,P\}} [f_{\sigma_t^p}(\boldsymbol{x}_d, \mathbf{\Theta}_t^p)]. \tag{6}$$

We then refine the shape network weights σ_t^p of the avatars by minimizing the loss:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(\hat{o}_t, o_t^{raw}) + \lambda_{\text{coll}}\mathcal{L}_{\text{coll}} \tag{7}$$

where $\mathcal{L}_{\text{BCE}}(\hat{o}_t, o_t^{raw})$ is the binary cross entropy between the composited occupancy prediction and the corresponding point o_t^{raw} on the input scan. This encourages segmented avatars to, together, align well with the fused scan.

A key challenge is to correctly model contact-induced deformation. Recall that the avatars are initialized from non-contact frames and hence do not yet account for the flattening of clothing and soft tissue due to contact. Therefore, the pose optimization step will cause surfaces that are in contact to intersect. To alleviate this, we select points that are predicted to be inside multiple subjects by querying the individual occupancies. We denote this subset of points as $\mathcal{S} = \{\boldsymbol{x}_d \mid \hat{o}_t^i(\boldsymbol{x}_d) > 0.5, \hat{o}_t^j(\boldsymbol{x}_d) > 0.5 \; \forall \; i \in \{1, \ldots, P\}, j \in \{1, \ldots, P\}, i \neq j\}$. We then penalize interpenetration of surfaces via $\mathcal{L}_{\text{coll}}$:

$$\mathcal{L}_{\text{coll}} = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}_d \in \mathcal{S}} \phi(\hat{o}_t^i(\boldsymbol{x}_d)) \cdot \phi(\hat{o}_t^j(\boldsymbol{x}_d)), \tag{8}$$

where $\phi(\hat{o}_t(\boldsymbol{x}_d)) = \max(\hat{o}_t(\boldsymbol{x}_d) - 0.5, 0)$. Intuitively, we ask that the uniformly sampled 3D points do not result in occupancy values of 1 (i.e., inside) for multiple shapes simultaneously. Instead, the shape networks are optimized such that the surfaces adhere to contact deformation.

6. Dataset

We believe that with Hi4D we contribute a valuable tool for the community working on human-to-human close interaction. For a detailed list of its contents and comparison to existing datasets, please refer to Tab. 1 and Supp. Mat.

We recruited 20 unique pairs of participants with varying body shapes and clothing styles to perform diverse interaction motion sequences of a few seconds, such as hugging, posing, dancing, and playing sports. We collected 100 independent clips with 11K frames in total and more than 6K frames of them with physical contact. Contact annotations in Hi4D cover over 95 % of the parametric human body. The contact coverage and contact frequency of Hi4D is shown in Fig. 4.

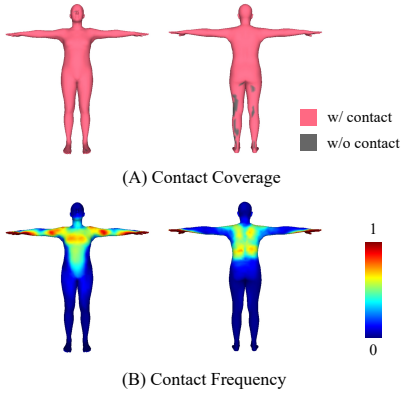

Figure 4. **Contact coverage and frequency of Hi4D.**

7. Experiment

We conduct ablations on Hi4D to verify our design choices (Sec. 7.1) and to compare with baselines (Sec. 7.2). We consider the following metrics for reconstruction evaluation: volumetric IoU (mIoU) [%], Chamfer distance (C-L_2) [cm], point-to-surface distance (P2S) [cm], and normal consistency (NC) [%].

Method	IoU ↑	C-L_2 ↓	P2S ↓	NC ↑
Ours (w/o shape refine)	0.982	0.36	0.37	0.934
Ours (w/o alternating opt.)	0.938	0.48	0.46	0.927
SMPL+D	0.983	0.24	0.30	0.927
Ours	**0.989**	**0.22**	**0.23**	**0.945**

Table 2. **Quantitative Results.** Ablations to evaluate our method without the shape refinement stage and without alternating optimization and comparison to the SMPL+D baseline.

7.1. Ablation Study

Shape Refinement. We compare our full optimization pipeline to a version without the shape refinement stage (Sec. 5.3). From Fig. 5, we observe that the details in the cloth are not accurately modelled if we do not further refine the shape network weights. Moreover, without the shape refinement stage, the method fails to model the contact-aware cloth deformations (*e.g.* the right hand of the blue colored person is occluded by the other person's cloth). Note that such deformations only occur with physical contact between people and cannot be learned from individual scans. The quantitative results in Tab. 2 also support the benefits of the shape network refinement.

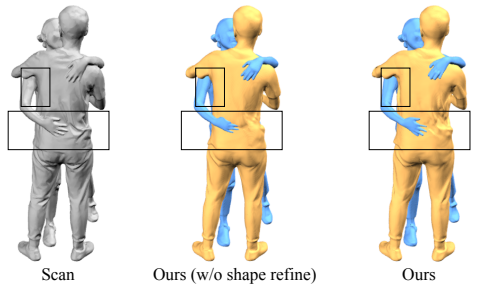

Figure 5. **Qualitative ablation (shape refinement).** The shape refinement stage better models the contact-aware deformation.

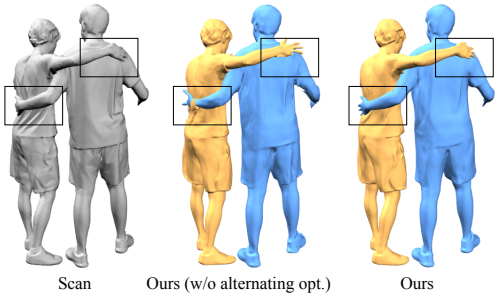

Figure 6. **Qualitative ablation (alternating optimization).** Optimizing poses and shape networks alternatingly improves results in areas with heavy contact.

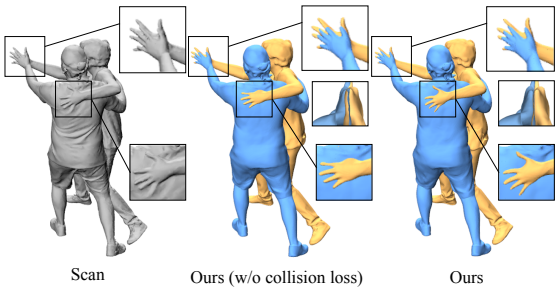

Figure 7. **Importance of collision loss.** Instance meshes intersect each other in contact areas if we remove the collision loss term.

Alternating Optimization. We compare our alternating optimization with an approach where poses and shape networks are optimized concurrently. We observe that in this case, it is hard to disambiguate body parts of different sub-

Setting	Method	MPJPE ↓	MVE ↓	NMJE ↓	NMVE ↓	F1 ↑	PCDR$^{0.10}$ ↑	CD ↓
Monocular	PARE [42]	87.6	106.5	95.3	115.9	0.919	0.610	297.7
	ROMP [66]	93.0	116.2	93.2	116.4	0.998	0.613	338.0
	BEV [67]	92.5	113.7	92.6	113.8	0.999	0.745	295.8
Multi-view	MVPose (4-views) [18]	61.3	78.3	67.0	85.4	0.917	0.957	234.8
	MVPose (8-views) [18]	50.3	61.8	51.8	63.6	0.971	0.972	166.8

Table 3. **SMPL estimation.** Results of monocular and multi-view SMPL estimation methods on Hi4D (*cf*. Sec. 8.1 and Fig. 8).

Setting	Method	IoU ↑	C-L_2 ↓	P2S ↓	NC ↑
Monocular	PIFuHD [61]	0.761	3.02	2.89	0.755
	ICON [73]	0.780	2.76	2.54	0.762
Multi-view	DMC [77] (4-views)	0.893	1.78	1.82	0.832
	DMC [77] (8-views)	0.906	1.64	1.60	0.851

Table 4. **Detailed geometry reconstruction.** Results of monocular and multi-view detailed geometry reconstruction methods on Hi4D (*cf*. Sec. 8.2 and Fig. 9).

jects in the contact area (*cf*. Fig. 6). Our alternating pipeline can better disentangle the effects of the pose and shape network. The quantitative results of Tab. 2 also confirm this.

Collision Loss. Without penalizing the collision of the two occupancy fields, one person's mesh might be partially intersected by the other person in the contact area, as we see from Fig. 7. We quantitatively measure the interpenetration by calculating the intersection volume between the individual segmented meshes. With the collision loss term defined in Eq. (8), the average intersection volume decreases from $11.62 \times 10^{-4}\,\mathrm{m}^3$ to $5.82 \times 10^{-4}\,\mathrm{m}^3$ by 49.91%.

7.2. Comparison Study

SMPL+D Baseline. A straightforward baseline for our task is to directly track multiple clothed SMPL body template meshes. We define this baseline as the SMPL+D (*cf*. [4, 5]) tracking baseline. It tracks the 3D geometry of close interacting people at each frame by estimating the individual displacement of each SMPL vertex and each subject. We optimize these displacement fields and the SMPL body parameters in a similar alternating manner as we do in our proposed method. For more implementation details please refer to the Supp. Mat. Quantitatively, our proposed method with personalized priors outperforms the SMPL+D baseline on all metrics (Tab. 2). The optimized SMPL+D models do not have any personalized prior knowledge from which we can infer when substantial instance ambiguity exists. We show qualitative results in the Supp. Mat.

8. Benchmark Baselines

We define several standard vision benchmarks conducted on the Hi4D dataset. These benchmarks include monocular SMPL estimation, multi-view SMPL estimation, monocular detailed geometry reconstruction and multi-view detailed geometry reconstruction. We evaluate several baseline methods on each of these tasks and demonstrate experimentally that our Hi4D dataset is challenging, thus opening many doors for future research.

8.1. SMPL Estimation

Evaluation Protocol. Hi4D provides multi-view RGB sequences with corresponding SMPL body registrations of the interacting people. For the SMPL estimation task, we mainly follow the evaluation protocol of AGORA [56]. For the monocular setting, MPJPE [mm] and MVE [mm] are calculated after alignment to the pelvis. The NMJE [mm] and NMVE [mm] are MPJPE and MVE errors normalized by the F1 score respectively. We adopt the Percentage of Correct Depth Relations (PCDR$^{0.1}$ [%]) metric that is introduced in [67] to evaluate depth reasoning. Contact Distances (CD [mm]) measures the distances between contact correspondences annotated in our dataset.

Monocular Setting. We evaluate one top-down method (PARE [42]) and two bottom-up methods (ROMP [66] and BEV [67]) for the monocular SMPL estimation task. From Tab. 3 we see that all methods have a relatively high MPJPE and MVE, demonstrating that current methods are not robust enough when strong human-human occlusion occurs. All methods fail to provide the reasonable spatial arrangement and contact relation, as shown in metric CD and Fig. 8.

Multi-View Setting. Most of the multi-view pose estimation methods still focus only on skeleton estimation without taking body shape into account. We evaluate the open-sourced multi-view SMPL estimation method MVPose [18] on 4-view and 8-view settings. Although in the multi-view setting, MVPose achieves lower MPJPE and MVE, heavy interpenetration, and inaccurate poses in 3D space still exist, especially in the contact area (*cf*. Fig. 8).

8.2. Detailed Geometry Reconstruction

Evaluation Protocol. Hi4D provides high-quality 4D scans, which can serve as ground truth for the detailed geometry reconstruction task. We apply the same metrics described in Sec. 7 to measure the reconstruction accuracy.

Monocular Setting. Most of the existing monocular mesh reconstruction methods focus on the single-person scenario without any occlusion. We extend two methods for single-person geometry reconstruction PIFuHD [61] and ICON [73] to handle the multi-person case. For implementation

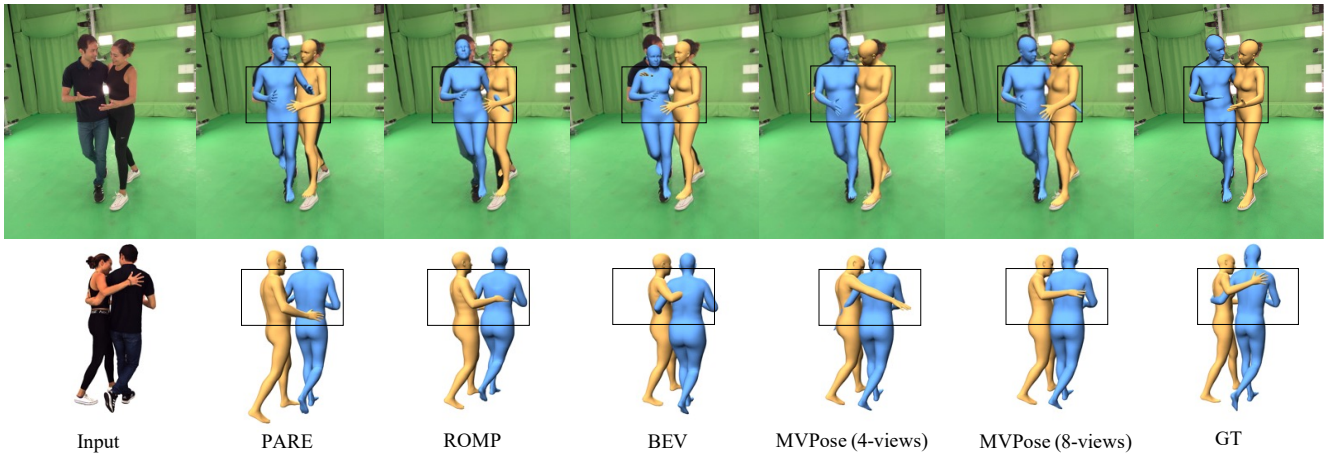

Figure 8. **SMPL estimation.** First row: results in 2D image space. Second row: results in an alternative 3D view (*cf*. Sec. 8.1).

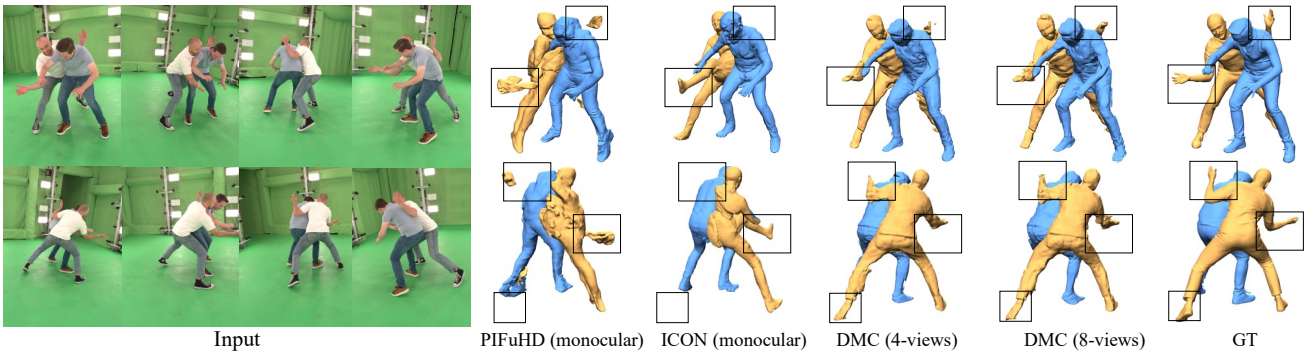

Figure 9. **Detailed geometry reconstruction.** Results of monocular and multi-view methods together with the GT of Hi4D (*cf*. Sec. 8.2).

details please refer to the Supp. Mat. From Fig. 9, we can observe that both methods are not robust against human-human occlusions and fail to produce high-quality reconstructions. Tab. 4 quantitatively shows that current single-person methods cannot achieve satisfactory reconstructions when directly extended to the challenging multi-person scenario. We believe that Hi4D provides the necessary data to unlock next-generation methods to reconstruct detailed geometry from monocular RGB sequences depicting closely interacting humans.

Multi-View Setting. We evaluate the method DMC [77] in both the 4-view and 8-view settings. Qualitative results in Fig. 9 show that although DMC can correctly reconstruct the geometry globally, artifacts still exist, *cf*. the hands and feet of the person colored in yellow. Tab. 4 further highlights the opportunities for improvement on this task.

9. Conclusion

In this paper, we propose a method to track and segment 4D scans of multiple people interacting in close range with dynamic physical contact. To do so we first build a personalized implicit avatar model for each subject and then refine pose and shape network parameters given fused raw scans in an alternating fashion. We further introduce Hi4D, a dataset consisting of close human interaction with high-quality 4D textured scans alongside corresponding multi-view RGB sequences, instance segmentation masks in 2D and 3D, registered parametric body models and vertex-level contact annotations. We define several vision benchmarks, such as monocular and multi-view human pose estimation and detailed geometry reconstruction conducted on Hi4D.

Limitations. Currently, our method does not model hands or facial expressions explicitly. We see the integration of more expressive human models *e.g*. [64] as a fruitful future direction. Furthermore, the optimization schema of our method is not very computationally efficient. The optimization can be accelerated remarkably by upgrading the current deformer to a faster version [12].

Acknowledgments. We thank Stefan Walter and Dean Bakker for the infrastructure support. We thank Deniz Yildiz and Laura Wülfroth for the data collection. We also thank all the participants who contribute to Hi4D.

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 2

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, jul 2005. 2

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 4, 7

[5] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Advances in Neural Information Processing Systems (NeurIPS)*, December 2020. 4, 7

[6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct 2019. 2

[7] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022. 3

[8] Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2300–2308, 2015. 2

[9] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 4, 5

[10] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 2

[11] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 4

[12] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-SNARF: A fast deformer for articulated neural fields. *arXiv*, abs/2211.15601, 2022. 8

[13] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 2, 3, 4

[14] Yin Chen, Z. Cheng, Chao Lai, Ralph Robert Martin, and Gang Dang. Realtime reconstruction of an animating human body from a single depth camera. *IEEE Transactions on Visualization and Computer Graphics*, 22:2000–2011, 2016. 2

[15] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[16] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2

[17] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *European Conference on Computer Vision*, pages 612–628. Springer, 2020. 2, 4

[18] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7792–7801, 2019. 3, 7

[19] Zijian Dong, Jie Song, Xu Chen, Chen Guo, and Otmar Hilliges. Shape-aware multi-person pose estimation from multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11158–11168, 2021. 3

[20] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3D-MPA: Multi Proposal Aggregation for 3D Semantic Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[21] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[22] 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. 1, 2, 3

[23] Coert van Gemeren, Ronald Poppe, and Remco C Veltkamp. Spatio-temporal detection of fine-grained dyadic human interactions. In *International Workshop on Human Behavior Understanding*, pages 116–133. Springer, 2016. 2, 3

[24] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2

[25] 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

[26] 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. *arXiv*, 2023. 3

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

[28] 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. 2

[29] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4421–4430, 2019. 2

[30] Tao Hu, Xinyan Zhu, Wei Guo, and Kehua Su. Efficient interaction recognition through positive action representation. *Mathematical Problems in Engineering*, 2013, 2013. 3

[31] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, June 2022. 3

[32] 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*, pages 3093–3102, 2020. 3

[33] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 3

[34] 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. 3

[35] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 3

[36] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. 2020. 3

[37] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[38] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2, 3

[39] Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Mask transfiner for high-quality instance segmentation. In *CVPR*, 2022. 2

[40] 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. 2

[41] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3

[42] 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*, pages 11127–11137, 2021. 3, 7

[43] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 2, 3

[44] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 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*, pages 3383–3393, 2021. 2, 3

[45] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3132–3141, 2022. 2

[46] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2720–2735, 2013. 2

[47] Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR 2011*, pages 1249–1256. Ieee, 2011. 2

[48] Yong Liu, Rangming Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. In *ECCV*, 2022. 2

[49] Matthew Loper, Naureen Mahmood, and Michael J. Black. Mosh: Motion and shape capture from sparse markers. *ACM Trans. Graph.*, 33(6), nov 2014. 2

[50] 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. 2, 4

[51] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3d people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[52] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of

motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 2

[53] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 2, 3

[54] 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

[55] 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. 3

[56] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 7

[57] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2

[58] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[59] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 91–99, Cambridge, MA, USA, 2015. MIT Press. 2

[60] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, 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. 3

[61] 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. 3, 7

[62] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 2, 4

[63] Hongje Seong, Seoung Wug Oh, Brian L. Price, Euntai Kim, and Joon-Young Lee. One-trimap video matting. In *ECCV*, 2022. 2

[64] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[65] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision*, pages 744–760. Springer, 2020. 3

[66] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021. 2, 3, 7

[67] 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*, pages 13243–13252, 2022. 3, 7

[68] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Human instance matting via mutual guidance and multi-instance refinement. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2637–2646, 2022. 2

[69] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[70] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11708–11718, 2021. 2, 4

[71] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct multi-view multi-person 3d human pose estimation. *Advances in Neural Information Processing Systems*, 2021. 3

[72] Guo Wen, Bie Xiaoyu, Alameda-Pineda Xavier, and Moreno-Noguer Francesc. Multi-person extreme motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3

[73] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 3, 7

[74] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[75] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, June 2018. 2

[76] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Lightweight multi-person total motion capture using sparse multi-view cameras. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 5560–5569, 2021. 3

[77] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6239–6249, 2021. 2, 3, 7, 8

[78] 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*, 44(6):3170–3184, 2021. 3