Hi4D: 4D Instance Segmentation of Close Human Interaction

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

Contents

1. Challenges 1

2. Dynamic Personalized Prior 1
   2.1. SMPL Registration 1
   2.2. Network Architecture 1

3. Instance Segmentation During Interaction 2
   3.1. Data Preprocessing 2
   3.2. Implementation Details 2

4. Hi4D Dataset 3
   4.1. Capture System 3
   4.2. Contents 3
   4.3. Subject Statistics 3
   4.4. Pose Accuracy Validation 3
   4.5. Ethics 4

5. Experiments 4
   5.1. Number of Alternating Optimization Steps 4
   5.2. SMPL+D Baseline 4
   5.3. SNARF (w/o pre-built dynamic personalized priors) Baseline 5
   5.4. Results on More Than Two People 5

6. Benchmark 5
   6.1. SMPL Estimation 5
   6.2. Detailed Geometry Reconstruction 6
   6.3. Additional Notes 6

7. Societal Impact 6

1. Challenges

Although our multi-view, volumetric capture systems can provide high-quality textured scans of individuals, it fuses the 3D surfaces of multiple closely interacting persons into a single connected surface. In Fig. 10 we show several examples of the fused geometry of multiple persons interacting with physical contact. As we can see from Fig. 10, the raw scan does not contain any instance-level information, thus there exists a lot of instance ambiguity in the contact area. Our main challenges are to derive complete per-subject surface geometry from the fused scan and to further gain instance-level information in 3D space.

2. Dynamic Personalized Prior

2.1. SMPL Registration

Registering the SMPL model [13] to individual scans is formulated as an energy minimization problem over body shape $\beta$, pose $\theta$ and translation $t$ as defined in Eq. (1) in the main manuscript. The important energy terms are detailed as following:

**Surface Energy Term** $E_S$: bi-directional Chamfer distance between the scan $R$ and registered SMPL template $M$ defined by

$$E_S = \frac{1}{|V_M|} \sum_{v_M \in V_M} \min_{v_R \in V_R} \rho\left(\|v_M - v_R\|\right) + \frac{1}{|V_R|} \sum_{v_R \in V_R} \min_{v_M \in V_M} \rho\left(\|v_M - v_R\|\right),$$

where $V_R$ and $V_M$ are the vertices of the raw scan and the SMPL template, respectively. $\rho$ is the Geman-McClure robust penalty function.

**3D Keypoint Energy Term** $E_J$: we first detect the 2D keypoints on the multi-view images via [3]. The 3D keypoints $J_{3D}$ are then obtained via robust triangulation of the 2D keypoints. The keypoint energy term is then formulated as

$$E_J = \frac{1}{|J|} \sum_{j=1}^{J} \|J_{\text{SMPL}}(\theta, \beta, t)_j - J_{3D_j}\|,$$

where $J_{\text{SMPL}}(\theta, \beta, t)$ is the 3D SMPL joints given the SMPL parameters.

Several examples of SMPL registrations to individual scans are shown in Fig. 11.

2.2. Network Architecture

We follow [4] to use two neural networks to model shape and deformation in canonical space. The network architec-
The raw scans fuse the individual surface geometries into a single connected geometry and thus do not contain any instance-level information.

Figure 11. Examples of SMPL registrations.

The shape field and deformation field are illustrated in Fig. 12. To better model the high-frequency details such as wrinkles of clothed humans, positional encoding [14] with 4 frequency components is applied to the input points.

3. Instance Segmentation During Interaction

3.1. Data Preprocessing.

We capture each interaction sequence starting from a frame where no physical contact between the P subjects occurs. Note that such raw scans without physical contact can be easily decomposed into P connected components, i.e. the individual textured scans of the P subjects. We track the raw scans until the frame where the number of the decomposed component decreases, meaning there exists physical contact between subjects. We denote the last frame before the contact as $t_0$ and use the decomposed individual scans to obtain initial SMPL parameters $\Theta^0$.  

3.2. Implementation Details.

The scan-to-mesh loss term $L_{s2m}$ is generally defined as

$$L_{s2m}(S, M) = \frac{1}{|V_S|} \sum_{v_S \in V_S} \rho\left(\min_{v_M \in V_M} \|v_M - v_S\|\right), \quad (11)$$

where $\rho$ is the Geman-McClure robust penalty function.

We use the Adam optimizer [10] with the default values $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for the pose and shape optimization. In the pose optimization stage, the learning rate is set to $\eta = 10^{-2}$ and body poses are optimized until convergence. During the shape refinement stage, the learning rate is set to $\eta = 5 \times 10^{-5}$ and the weight of the collision loss term $\lambda_{coll}$ is 0.1. We observe that setting the number of alternating optimization steps $N$ to 2 can already lead to good convergence. For more discussion about the number of alternating optimization steps please refer to Sec. 5.1.
4. Hi4D Dataset

4.1. Capture System

We captured our dataset in a Volumetric Capture Studio equipped with 106 synchronized cameras (53 RGB and 53 IR cameras), from which we release 8 RGB images equally distributed on the external perimeter. The sequences are filmed at 12 MP, 30 FPS, and within an effective capture volume of 2.8 m in diameter and 3 m in height. Each frame consists of a mesh with 80K faces with an estimated average error of 1-2 mm, and a texture map of $4 \times 4$ MP resolution [5].

4.2. Contents

With Hi4D we publish the following data:

1. **4D textured scans.** High-quality textured scans obtained on our multi-view capture stage [5].

2. **Instance segmentation masks in 2D and 3D.** Given our method, the raw scans are then segmented automatically by assigning the label of the closest individual reposed avatar to each vertex. These 3D segmentation masks are then projected to multi-view RGB images to obtain 2D instance masks.

3. **Parametric body models.** As part of the outputs of the alternating optimization process, SMPL registrations of each individual are obtained along with the instance meshes by our proposed method.

4. **Vertex-level contact annotations.** For each vertex on the instance/SMPL mesh, we compute the point-to-surface distance to the mesh of another person. If the distance is lower than a threshold (1 cm for instance meshes and 2 cm for SMPL meshes) and the normals depict quasi-opposite direction, the vertex is labeled as in contact. In this way, we obtain a binary contact label for each vertex. We further find the contact correspondence of each in-contact vertex by searching for the closest contact point of another person. We denote the contact segmentation of a person $p$ as $S(p) \in \{0, 1\}^{N_{verts} \times 1}$ and the contact correspondence between person $p_0$ and $p_1$ as $C(p_0, p_1) \in \{0, 1\}^{N_{verts} \times N_{verts}}$.

5. **RGB images.** For every frame, we provide 8 RGB views as shown in Fig. 8 of the main paper.

More examples from Hi4D are shown in Fig. 21.

4.3. Subject Statistics

Hi4D captures 20 unique subject pairs (16 female, 24 male). Our dataset contains a variety of subject pairs with diverse height, weight and garments. The statistics of the participants are shown in Fig. 13.

4.4. Pose Accuracy Validation

Evaluating how accurate our SMPL registrations are is a challenging problem in itself. This is because we have frequent and heavy occlusions in our setting, so even the gold-standard, marker-based optical tracking, is struggling to produce accurate results without laborious manual interventions in post-processing. We take a first step towards evaluating the pose accuracy of our SMPL registrations by comparing our results to a capture technology that does not require line-of-sight but is still accurate. We chose to use electromagnetic (EM), body-worn sensors similar to [9].

More specifically, one subject is wearing 12 EM sensors, while performing 3 kinds of interactions with another subject on the volumetric capture stage. The EM sensors are synchronized with the cameras. We then fit the SMPL model to the EM data, assuming that the SMPL shape of the subject is known. Further, to spatially align the EM data with the coordinate frame of the capture stage, we track the EM source with an Apriltag [11, 16, 21]. We can then compare the SMPL fit obtained via the EM sensors with the SMPL fit obtained by our method. The results are shown in Tab. 5.

As we can see from Tab. 5, the error of our SMPL registration pipeline compared to EM-based fitting with body-
Table 5. Quantitative comparison to EM-based pose reference. Comparison of our SMPL registration pipeline to an SMPL fit obtained by fitting to body-worn EM sensors that do not require line-of-sight. We compare on three sequences and compute the average per-joint positional (MPJPE) and per-joint angular (MPJAE) error after Procrustes alignment.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>PA-MPJPE [mm]</th>
<th>PA-MPJAE [deg]</th>
</tr>
</thead>
<tbody>
<tr>
<td>dance</td>
<td>16.1</td>
<td>11.2</td>
</tr>
<tr>
<td>fight</td>
<td>18.5</td>
<td>9.1</td>
</tr>
<tr>
<td>hug</td>
<td>20.6</td>
<td>10.0</td>
</tr>
<tr>
<td>mean</td>
<td>18.4</td>
<td>10.1</td>
</tr>
</tbody>
</table>

Table 6. Ablation study on the number of alternating optimization steps.

<table>
<thead>
<tr>
<th>Number of Steps</th>
<th>IoU ↑</th>
<th>C-L₂ ↓</th>
<th>P2S ↓</th>
<th>NC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.987</td>
<td>0.23</td>
<td>0.23</td>
<td>0.945</td>
</tr>
<tr>
<td>2</td>
<td>0.989</td>
<td>0.20</td>
<td>0.21</td>
<td>0.946</td>
</tr>
<tr>
<td>5</td>
<td>0.991</td>
<td>0.19</td>
<td>0.20</td>
<td>0.947</td>
</tr>
</tbody>
</table>

4.5. Ethics

Our institution’s ethics committee duly approved the protocol we followed for the collection and publication of Hi4D. All subjects have freely volunteered to participate in this data collection. They have been duly informed about the intended use and publication of the dataset, signed a consent form, and have received compensation for the time it took to record them.

5. Experiments

5.1. Number of Alternating Optimization Steps

We select a subset of our collected data to evaluate the effect of the number of alternating optimization steps. With a larger number of alternating optimization steps, the reconstruction quality increases as shown in Tab. 6. The computational time increases proportionally to the number of optimization steps. In our implementation, the number of alternating optimization steps $N$ is set to 2 to balance between the reconstruction quality and the computational efficiency.
5.3. SNARF (w/o pre-built dynamic personalized priors) Baseline.

We further implement a baseline where instead of building avatars in advance we train the SNARF models of each subject jointly via the loss defined in Eq. 7 in the main manuscript. Note that training SNARF models from scratch requires accurate SMPL poses, which itself is a challenging problem especially when people interact in close proximity (see Sec. 8.1 in the main manuscript). In order to disentangle the influence of SMPL pose estimations, we use the reference SMPL pose obtained by our proposed method to build the avatars on the fly.

As seen from Fig. 16, without pre-built avatars, the results from joint training of multiple SNARF models from scratch tend to have artifacts, especially in the contact area. This observation further confirms the importance of creating individual avatars beforehand, which helps to tackle the instance ambiguity when multiple instances interact with physical contact. The quantitative results in Tab. 7 also verify that our proposed method can achieve better reconstruction accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU ↑</th>
<th>C-L₂ ↓</th>
<th>P2S ↓</th>
<th>NC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o pre-built priors</td>
<td>0.952</td>
<td>0.49</td>
<td>0.49</td>
<td>0.939</td>
</tr>
<tr>
<td>Ours</td>
<td>0.989</td>
<td>0.22</td>
<td>0.23</td>
<td>0.945</td>
</tr>
</tbody>
</table>

Table 7. Quantitative comparison with SNARF (w/o pre-built dynamic personalized priors). Our method with pre-built avatars consistently outperforms the SNARF baseline without pre-built dynamic personalized priors.

5.4. Results on More Than Two People

Our method is extendable to more than two people as shown in Fig. 17.

6. Benchmark

6.1. SMPL Estimation

Contact Distances (CD). This metric measures the average distances of annotated contact correspondences (cf. Sec. 4.2):

\[ CD = \frac{1}{N_{\text{SMPL}}} \sum_{(v_0,v_1) \in C(p_0,p_1)} \phi_D(v_0,v_1), \quad (12) \]

where \((v_0,v_1)\) is a pair of vertices in contact and \(\phi_D(v_0,v_1)\) is the euclidean distance between this contact correspondence.

Contact Optimization. From the results of SMPL estimation methods (cf. Tab. 3 and Fig. 8 in the main manuscript) we can observe common errors presented as the formats of incorrect spatial arrangement as well as strong interpenetration in 3D space. We hope our dataset can drive research on multi-person pose and shape estimation along with contact modeling.

To motivate research on contact modeling, we conduct an experiment on a subset of our collected data (around 3000 frames) to show the importance of contact. We use the SMPL outputs from ROMP [19] as our initialization. As we can see from Tab. 8, refining the SMPL outputs from ROMP solely with the 2D ground-truth keypoints via the 2D re-projection loss cannot fully alleviate the problem. Thus we add two additional contact-relevant losses:

1) Contact Segmentation Loss: We draw inspiration from [7] and define the contact segmentation value \(S_{\text{pred}}(p)_i\) at a vertex \(v_i\) of subject \(p\) is defined as follows:

\[ S_{\text{pred}}(p)_i = \min\left(\frac{0.02}{d_i}, 1.0\right), \quad (13) \]

where \(d_i\) denotes the minimal distance of vertex \(v_i\) to another person and 0.02 m is the contact threshold. The contact segmentation loss compares the current contact segmentation map \(S_{\text{pred}}\) with our annotations \(S_{\text{gt}}\) over \(N_{\text{SMPL}}\) all SMPL vertices for both subjects:

\[ L_s(p_0,p_1) = \frac{1}{2 \times N_{\text{SMPL}}} \sum_{p \in \{p_0,p_1\}} |S_{\text{pred}}(p) - S_{\text{gt}}(p)|. \quad (14) \]
Without explicitly taking contact information into account, there exist interpenetration and implausible poses.

2) Contact Distance Loss: We also minimize the contact distance loss which measures the distance of the paired in-contact vertices \((v_0, v_1)\) from subjects \((p_0, p_1)\) respectively. \(C(p_0, p_1)\) is denoted as the set of all \(N_c\) contact pairs.

\[
L_c(p_0, p_1) = \frac{\sum_{(v_0, v_1) \in C(p_0, p_1)} \phi_d(v_0, v_1)}{N_c}
\] (15)

From Tab. 8 we observe that the pose and shape estimation can further benefit from the correct contact modeling. A qualitative result can be found in Fig. 18.

<table>
<thead>
<tr>
<th>Method</th>
<th>MPIPE ↓</th>
<th>MVE ↓</th>
<th>PCDR↑↑</th>
<th>CD ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROMP</td>
<td>110.1</td>
<td>135.2</td>
<td>64.24</td>
<td>275.2</td>
</tr>
<tr>
<td>w/ 2D Kp</td>
<td>74.1</td>
<td>87.5</td>
<td>70.56</td>
<td>181.8</td>
</tr>
<tr>
<td>w/ 2D Kp + Contact</td>
<td>72.7</td>
<td>83.8</td>
<td>78.83</td>
<td>35.1</td>
</tr>
</tbody>
</table>

Table 8. Quantitative results of contact optimization on a subset of Hi4D.

6.2. Detailed Geometry Reconstruction

Monocular Setting. To our best knowledge, the only method that deals with multi-person reconstruction from a single image [15] does not handle the case where multiple people are interacting in close range and it is unfortunately not open-sourced. Thus we extend the single-person reconstruction methods PIFuHD [18] and ICON [22] to the multi-person case. More specifically, first, a pre-trained instance segmentation network [12] is applied to generate instance masks. The segmented images of each individual are given as input to PIFuHD [18] and ICON [22].

To evaluate the reconstruction performance, we first assign each predicted instance a ground truth instance ID by comparing the overlap region between predicted instance segmentation masks and ground truth instance segmentation masks. Then we perform ICP registration [1] between the reconstructed mesh (after scaling by height) and its corresponding ground truth mesh to align them in 3D space. After these processing steps, the reconstruction performance is evaluated with the same metrics mentioned in Sec. 7 of the main manuscript.

Multi-view Setting. Note that DMC [24] requires the SMPL-X models generated by [23], which are not publicly available. Instead, we use the output from MVPose [6] and convert the SMPL model to SMPL-X by using the official conversion tool [17].

More qualitative results of SMPL estimation and detailed geometry reconstruction methods are shown in Fig. 19 and Fig. 20.

6.3. Additional Notes

In the monocular setting, one camera view for each sequence is selected for evaluation. The information regarding the selected camera view will be released along with the dataset.

7. Societal Impact

Our dataset, Hi4D, promotes progress in 3D human pose and shape reconstruction from single or multiple RGB images. Such technology promises valuable applications that would benefit society at large, e.g., remote telepresence, automated rehabilitation, or computer-guided fitness and health coaches. However, human pose estimation, especially from images, might be abused for malicious surveillance or person identification via gait analysis or face recognition. Although neither our method nor our dataset directly caters to such dubious uses, it may foster future advancements of such methods and thus indirectly contribute to adverse uses. This poses an ethical and societal concern, which must be considered in future developments of these technologies. We argue that one way of doing so is to conduct transparent and open-sourced research to both inform the public about how such methods work exactly and to promote the research of respective countermeasures.
Figure 19. Qualitative results of SMPL estimation methods.

Figure 20. Qualitative results of detailed geometry reconstruction methods.
Figure 21. More examples from Hi4D.
References


