# Robust Human Registration with Body Part Segmentation on Noisy Point Clouds

Kai Lascheit, Marc Pollefeys, Daniel Barath
ETH Zurich, Switzerland

Leonidas Guibas, Francis Engelmann
Stanford University, USA

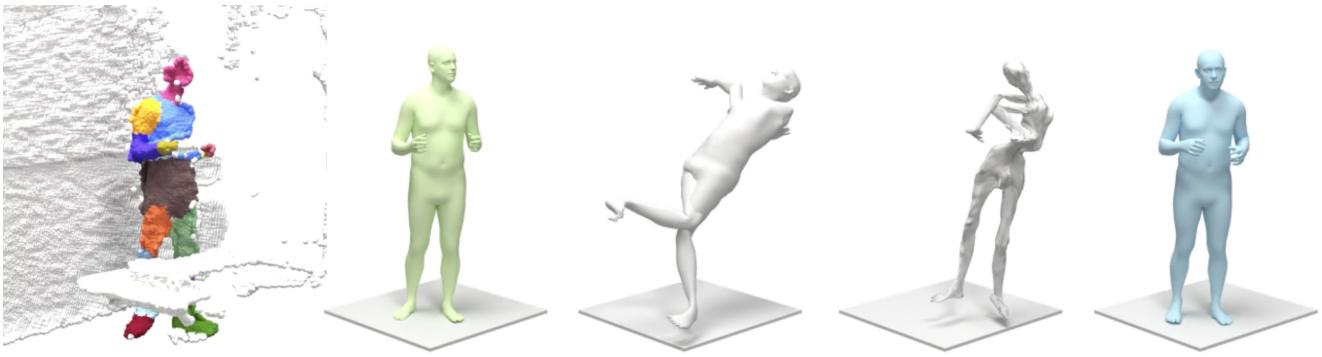


Figure 1. Our method SegFit reconstructs human poses from point clouds using body part segmentation and the SMPL-X model [25]. We showcase SMPL-X fitting results on the EgoBody dataset [38], and compare to the state-of-the-art methods ArtEq [12] and NICP [22].

## Abstract

*Registering human meshes to 3D point clouds is essential for applications such as augmented reality and human-robot interaction but often yields imprecise results due to noise and background clutter in real-world data. We introduce a hybrid approach that incorporates body part segmentation into the mesh fitting process, enhancing both human pose estimation and segmentation accuracy. Our method first assigns body part labels to individual points, which then guide a two-step SMPL-X fitting: initial pose and orientation estimation using body part centroids, followed by global refinement of the point cloud alignment. By leveraging semantic information from segmentation, our approach ensures robust alignment even with occlusions or missing data. Furthermore, it generalizes across diverse datasets without requiring extensive manual annotations or dataset-specific tuning. Additionally, we demonstrate that the fitted human mesh can refine body part labels, leading to improved segmentation. Evaluations on the challenging real-world datasets InterCap, EgoBody, and BEHAVE show that our approach significantly outperforms prior methods in both pose estimation and segmentation accuracy. Code and results are available on our project website: https://segfit.github.io.*

## 1. Introduction

Registering parametric human meshes to 3D point clouds involves estimating the human pose as well as shape parameters. Capturing the intricate details of human movement and shape is integral to many applications, ranging from creating realistic virtual avatars [28, 30] to improving human-robot interactions [2, 6]. Although fitting a parametric human body model to 3D data (*e.g.*, from LiDAR or Kinect) is a key step in such applications, the lack of contextual guidance often leads to imprecise results [7, 18]. By contrast, prior knowledge of body parts can greatly enhance fitting accuracy [32, 35, 37]. Motivated by this, we propose a novel method that jointly leverages body part segmentation and pose fitting on partial 3D point clouds.

In recent years, parametric models like SMPL-X [20] have become the de facto standard for describing 3D human poses and shapes. These models are typically fitted to point cloud data using either gradient-based optimization [7, 25] or neural networks [17, 18]. Parallel progress in body part segmentation – labeling each point in a cloud according to anatomical regions – has opened up new possibilities in computer graphics, healthcare, and autonomous systems [19, 29, 35]. However, current methods frequently underperform in real-world scenarios characterized by complex poses, occlusions, multi-person inter-

actions, and human-object interactions [14, 24, 31]. Many approaches are trained on synthetic datasets [1, 21], causing performance degradation when confronted with the variability of in-the-wild data [10, 23]. Additionally, these methods often struggle with generalization to unseen body poses and varying sensor noise, limiting their applicability in practical settings.

To address these limitations, we propose a hybrid framework that merges pose fitting and body part segmentation, allowing each to refine the other. We begin with an initial segmentation from the Human3D network [29], which provides a coarse assignment of points to body parts and is already fine-tuned on the in-the-wild EgoBody dataset [38]. This segmentation serves as the foundation for a two-step optimization procedure that fits the SMPL-X model. First, body part centroids guide an approximate alignment of the model pose and orientation [7, 34], establishing a robust initial configuration. Second, the model is refined by considering all points in the cloud, thereby capturing more nuanced pose and shape information [18, 25]. Throughout this process, we incorporate a pose prior [25] to ensure anatomically plausible body configurations, mitigating errors caused by occlusions or missing data. After fitting, we reassign body parts to the 3D point cloud via majority voting over nearest neighbors [13, 27], yielding a segmentation more accurate than the initial prediction. By uniting part segmentation and mesh fitting, our approach more effectively generalizes to diverse environments and remains robust under challenging conditions.

We evaluate our approach on three complex datasets – InterCap [15], EgoBody [38], and BEHAVE [5] – featuring occlusions, multiple interacting subjects, and human-object interactions [14, 26, 34]. Notably, we show strong performance on single-view depth data from InterCap and EgoBody, demonstrating that our method does not rely on multi-view setups. Compared to leading methods [12, 22, 29], we observe an up to *tenfold* boost in pose modeling accuracy and an up to *22%* gain in segmentation accuracy.

Our contributions are summarized as follows:

- **Segmentation-Based Pose Fitting.** A unified approach that integrates human pose fitting with body part segmentation on point clouds, increasing the accuracy and robustness of the fitting process in the presence of noisy point clouds.
- **Pose Fitting-Enhanced Segmentation.** We leverage fitted SMPL-X meshes to refine body part labels, resulting in more accurate segmentation. We also show how the more accurately segmented point clouds can be used to finetune a segmentation network self-supervised.

Our work advances the state of the art in 3D human body fitting and segmentation for point clouds, promising more faithful human representations in real-world, unstructured environments [6, 28, 30].

## 2. Related Work

Estimating human body pose and shape from 3D point clouds is vital in computer vision, with applications in virtual reality, animation, and human-computer interaction. While extensive research has been conducted on fitting parametric human models to 2D images [7, 16, 18], we focus on methods that directly operate on 3D point cloud data. Point clouds capture detailed geometric information and avoid the ambiguities inherent in 2D projections, making them valuable for precise human modeling.

**Human Registration on Point Clouds.** Several approaches have been developed to fit human poses to point clouds. Bhatnagar et al. [3] introduced *IP-Net*, which combines implicit representations with parametric models to reconstruct clothed human bodies from partial scans. IP-Net learns a continuous occupancy field representing the human body, allowing for detailed reconstructions even with incomplete data. Wang et al. [33] proposed *PTF*, a method that fits SMPL models to point clouds by considering local geometric features. By leveraging local point distributions, PTF improves fitting accuracy in areas with high curvature or fine-grained details. Zuo et al. [39] presented a self-supervised approach for 3D human motion reconstruction from depth sequences. Their method leverages temporal coherence without requiring ground truth annotations, effectively reconstructing dynamic human motions. Cai et al. [9] developed *PointHPS*, a hierarchical point-based network that directly regresses SMPL parameters from point clouds. PointHPS achieves state-of-the-art results using a point-based encoder-decoder architecture.

Recently, Marin et al. [22] introduced *NICP (Neural ICP)*, which bridges classic ICP with learnable shape representations. NICP iteratively refines a neural deformation field at inference time to better align human models with input point clouds, demonstrating notable improvements in registration accuracy across challenging poses and noisy scans. Moreover, Feng et al. [12] proposed *ArtEq*, a part-based SE(3)-equivariant neural network for SMPL fitting. Unlike previous learning-based methods that struggle with out-of-distribution poses, ArtEq explicitly incorporates SE(3) invariance and equivariance to improve generalization. The method achieves state-of-the-art accuracy on the PosePrior subset of AMASS [21] and is significantly faster than previous methods.

Despite these advances, common challenges persist. Methods often rely on large annotated datasets or specialized training for generalization, which may be challenging to scale. Regression-based approaches can miss subtle body shape variations or fail under significant occlusions. Optimization-based algorithms are sensitive to initialization and can get stuck in local minima, especially for complex

*Body-Parts* *Centroids* *Segmentation of two humans*

Figure 2. The 15 body parts and their centroids *(left)*. Example body parts segmentation of two humans from our SegFit *(right)*.

poses with self-contact or multiple people interacting. Our work addresses these challenges by integrating body part segmentation into the fitting process, leveraging semantic cues to distinguish between symmetric limbs and reduce orientation ambiguities. The segmentation-informed initialization obviates the need for multiple optimization runs (as in SMPLify-X [25]), ensuring more robust convergence without exhaustive restarts.

## 3. Method

We propose a framework for fitting an SMPL-X parametric body model to 3D point clouds by leveraging initial body part segmentation. Our method is built around two key processes: a robust model initialization guided by body part centroids and a subsequent iterative fitting phase that refines pose and shape parameters. Subsequently, we use the created fit to refine the initial segmentation based on a nearest-neighbor approach.

### 3.1. Problem Definition

Given a 3D point cloud $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N}$ representing one or more human subjects, we aim to estimate the pose and shape parameters of the SMPL-X model [25] such that the resulting 3D meshes accurately align with the underlying geometry. A primary challenge arises from the symmetry of the human body: without suitable initialization, standard optimization often converges to suboptimal minima. To address this, our method exploits initial body part segmentation from Human3D [29], thereby providing critical semantic guidance for both initialization and optimization.

### 3.2. Overview of the Approach

Our approach comprises five major steps:

1. **Initial Body Part Segmentation:** Use Human3D [29] to label each point in $\mathcal{P}$ with a body part index, producing a coarse yet valuable partitioning of the point cloud.
2. **Model Initialization:** Compute body part centroids (see Figure 2) from the segmentation and align these centroids with those of an SMPL-X template to establish a robust initial pose and orientation.
3. **Model Fitting:** Refine the SMPL-X parameters by optimizing a multi-term objective that balances data fidelity and pose/shape regularization over the entire point cloud.
4. **Enhanced Part Segmentation:** Assign body part labels via nearest-neighbor majority voting on the fitted mesh, generating a more accurate segmentation than the initial network output.

Below, we describe each component in detail.

### 3.3. Initial Body Part Segmentation

We begin by segmenting the input point cloud $\mathcal{P}$ with Human3D [29], a state-of-the-art network that predicts a body part label $s_i \in \{1, \ldots, K\}$ for every point $\mathbf{p}_i \in \mathcal{P}$. This step encodes high-level semantic cues about the spatial organization of the human body in 3D, enabling subsequent stages to distinguish between symmetric limbs and reduce ambiguity in pose initialization.

### 3.4. Model Initialization

A common pitfall in human model fitting is suboptimal initialization, which can hinder or derail convergence [14, 25]. Rather than performing multiple fitting trials with varied initial orientations, we leverage the body part segmentation to establish a direct, data-driven initialization. Specifically:

1. Compute centroid $\mathbf{c}_k^{\text{scan}} = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{p}_i$ for each body part $k$, where $N_k$ is the number of points labeled $k$.
2. Identify the corresponding centroids in the SMPL-X template $\mathcal{M}_0$, denoted by $\mathbf{c}_k^{\text{model}}$.
3. Compute a global rotation $\mathbf{R}_0$ and translation $\mathbf{t}_0$ that align $\mathbf{c}_k^{\text{model}}$ to $\mathbf{c}_k^{\text{scan}}$, providing a well-informed initial pose for the subsequent optimization.

This centroid-based matching approach effectively addresses orientation ambiguities by exploiting structural cues in the data, similar to finding the corner pieces of a puzzle before refining the interior.

### 3.5. Model Fitting

**SMPL-X Parameterization.** The SMPL-X model [25] is parameterized by pose $\boldsymbol{\theta} \in \mathbb{R}^{3J}$, shape $\boldsymbol{\beta} \in \mathbb{R}^{B}$, and global translation $\mathbf{t} \in \mathbb{R}^3$, where $J$ is the number of joints and $B$ is the dimension of the shape space. To avoid implausible configurations, we employ VPoser [25], a learned human pose prior that maps $\boldsymbol{\theta}$ to a latent space with higher-level constraints on body articulation.

**VPoser Prior.** VPoser [25] is a variational autoencoder (VAE)-based prior for human body pose which has been trained on the AMASS [21] dataset, capturing a large range of natural human poses. It applies an encoder-decoder architecture to transform SMPL-X parameters into a compact latent space. We optimize the SMPL-X parameters in

this latent space to constrain our method to realistic human poses, as is common in optimization approaches to human mesh fitting [14, 25].

**Objective Function.** We refine $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and $\mathbf{t}$ by minimizing a combined energy:

$$\mathcal{L} = \lambda_{\text{data}}\mathcal{L}_{\text{data}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{shape}}\mathcal{L}_{\text{shape}}, \quad (1)$$

with hyperparameters $\lambda_{\text{data}}$=1, $\lambda_{\text{pose}}$=0.5, $\lambda_{\text{shape}}$=0.5. A parameter study is presented in the experiments (Sec. 4.4). Below, we provide a description for each term.

*Data Term* quantifies alignment between the model surface and $\mathcal{P}$. We adopt a robust, one-sided Chamfer distance with a Huber loss to make it robust as follows:

$$\mathcal{L}_{\text{data}} = \sum_{k=1}^{K}\sum_{i=1}^{N_k} \min_{\mathbf{v}\in\mathcal{V}_k} \mathcal{L}_{\text{Huber}}(\mathbf{p}_i - \mathbf{v}), \quad (2)$$

where $\mathcal{V}_k$ denotes vertices belonging to body part $k$. A one-sided formulation prioritizes model-to-data consistency and mitigates erroneous penalization of regions without corresponding sensor capture.

*Pose Term* regularizes poses around the VPoser prior, encouraging kinematically realistic articulation:

$$\mathcal{L}_{\text{pose}} = \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2, \quad (3)$$

where $\boldsymbol{\theta}_0$ is the default pose of the SMPL-X model.

*Shape Term* constrains shape coefficients to reasonable magnitudes to describe a realistic human shape as follows:

$$\mathcal{L}_{\text{shape}} = \|\boldsymbol{\beta}\|_2^2. \quad (4)$$

We employ the Adam optimizer with early stopping (200 maximum steps) to minimize Eq. (1). This yields a refined pose and shape that closely aligns the SMPL-X body model to the input data.

### 3.6. Enhanced Part Segmentation

After fitting, we reassign part labels to each point $\mathbf{p}_i$ using majority voting among its nearest-neighbor vertices on the SMPL-X mesh. Formally, let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be the $n$ closest mesh vertices to $\mathbf{p}_i$, each labeled by a body part index. We obtain the new label $s_i$ via:

$$s_i = \arg\max_{k\in\{1,\ldots,K\}} \sum_{j=1}^{n} \alpha_j \, \delta\big(\text{label}(\mathbf{v}_j) = k\big), \quad (5)$$

where $\alpha_j$ is an inverse distance weight, and $\delta(\cdot)$ is the Kronecker delta. This label reassignment leverages the accurate surface alignment from the fitted model, producing a more reliable body part segmentation than the initial one.

### 3.7. Summary

By integrating pose fitting with segmentation, our method achieves robust initializations and more precise model alignments. This synergy surpasses approaches that treat pose estimation and segmentation separately, particularly in real-world scenarios where occlusions, noise, and high variability challenge purely data-driven or optimization-based techniques.

## 4. Experiments

In this section, we first compare our approach SegFit with state-of-the-art human registration methods on three challenging real-world datasets (Sec. 4.1). We then show how SegFit also improves 3D human body-part segmentation (Sec. 4.2). Next, we provide detailed analysis experiments to understand the importance of human body part segments for human pose estimation, and the effects of varying optimization strategies (Sec. 4.4). Finally, we show qualitative results of SegFit on *in-the-wild* environments (Sec. 4.5).

### 4.1. Comparing with State-of-the-Art Methods

**Datasets.** To evaluate the robustness of our approach, we rely on three challenging "in-the-wild" datasets [5, 15, 38], which feature cluttered scenes, occlusions, and noisy, partial observations (see Fig. 3). These datasets provide a more realistic setting compared to controlled "in-the-lab" datasets [8, 21, 36], where humans are isolated from the background and captured with high-quality cameras from multiple-view points in empty scenes.

*EgoBody* [38] is a large-scale dataset capturing ground-truth 3D human motions during social interactions in natural 3D environments. It includes up to two individuals interacting with each other and their surroundings, resulting in significant occlusions, partial observations, and challenges in disentangling humans from the background. *BEHAVE* [5] is a full-body human-object interaction dataset containing multi-view RGB-D sequences and annotated human meshes in natural environments. It features a single person interacting closely with various objects, leading to challenging poses and strong occlusions. *InterCap* [15] is another large-scale dataset focusing on human-object interactions, similar to BEHAVE. Each scene involves a single human interacting with one out of ten different object types. All three datasets are captured with multi-view Kinect RGB-D sensors which simplifies the accurate ground truth human mesh annotation. However, our experiments rely solely on single-view depth frames – more representative of real-world applications – leading to significant occlusions from objects in the scene and partial human observations.

**Baseline Methods.** We compare our SegFit with the most recent state-of-the-art methods for registering SMPL human meshes to 3D point clouds: *ArtEq* [12] introduces articu-

| Method | BEHAVE [5] V2V in mm | BEHAVE [5] MPJPE in mm | BEHAVE [5] Time in s | EgoBody [38] V2V in mm | EgoBody [38] MPJPE in mm | EgoBody [38] Time in s | InterCap [15] V2V in mm | InterCap [15] MPJPE in mm | InterCap [15] Time in s |
|---|---|---|---|---|---|---|---|---|---|
| ArtEq [12] | 140.6 | 162.4 | **0.105** | 538.7 | 605.7 | **0.098** | 422.0 | 515.0 | **0.103** |
| NICP [22] | 59.9 | – | 33.14 | 232.1 | – | 17.2 | 257.7 | – | 22.2 |
| SegFit (Ours) | **37.0** | **30.7** | 1.86 | **47.9** | **42.2** | 1.79 | **147.2** | **140.7** | 1.37 |

Table 1. Pose and shape scores of SegFit in comparison to NICP [22] and ArtEq [12] on the BEHAVE [5], EgoBody [38], and InterCap [15] datasets. Metrics are vertex-to-vertex (V2V) distance, mean-per-joint-position-error (MPJPE) and average runtime per-human.

lated SE(3)-equivariance for SMPL model fitting, enabling generalization to unseen poses by learning part-based transformations instead of global ones. It combines SO(3)-invariant part detection with pose- and shape-equivariant regression, leveraging self-attention layers to preserve equivariance. *NICP* [22] is a ICP-style self-supervised approach tailored to neural fields, enabling robust and scalable 3D human registration across diverse shapes and datasets without requiring manual annotations improving over comparable prior approaches [3, 4, 33].

To ensure a fair comparison, we apply one adjustment: since ArtEq and NICP are not designed for multi-human or cluttered scenes with backgrounds and occlusions, we first isolate human instances using Human3D's human instance segmentation results before processing the resulting human point clouds with NICP and ArtEq. This ensures that all methods process only points that belong to humans as predicted by Human3D [29].

**Metrics.** We follow prior work [12, 33] to evaluate the accuracy of the registered SMPL [20] human model. Shape error is measured as the Euclidean distance between corresponding mesh vertices, *i.e.*, the vertex-to-vertex (V2V) error in mm. Pose accuracy is evaluated as the mean per joint position error (MPJPE) in mm between the fitted and ground-truth SMPL models. We also report the average processing time required to fit a single human instance. For body-part segmentation, we follow [12, 29] and report accuracy (Acc), intersection over union (IoU), and mean average precision (mAP).


Figure 3. **Datasets.** Example scenes from BEHAVE [5], EgoBody [38], and InterCap [15]. We show RGB images for illustration only, all experiments are performed on single-view depth maps.

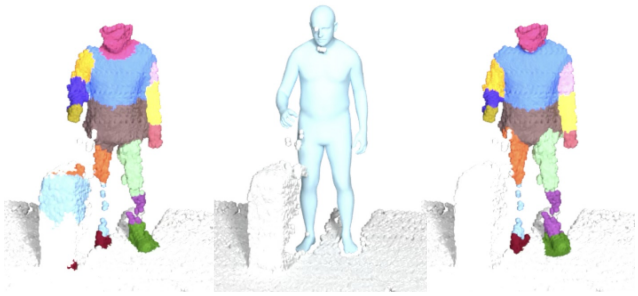


Figure 4. **Refined Body-Part Segmentation.** Example output of our SegFit on the InterCap [15] dataset. We show the initial body-part segmentation from Human3D [29] *(left)*, our registered human mesh *(center)*, and the improved human body-part segmentation based on nearest neighbors majority voting *(right)*. Notice that, at first, the suitecase is mistakenly labeled as the left leg. However, despite the occlusion, our approach successfully corrects the human pose and refines the body-part segmentation.

**Results.** Table 1 presents the results for human pose and shape fitting across all evaluated datasets. MPJPEs are not reported for NICP, as this method only predicts SMPL vertex positions. SegFit consistently outperforms all prior methods by a substantial margin, demonstrating superior generalization to a wide range of diverse and noisy real-world point clouds. Overall, SegFit achieves the highest performance on BEHAVE, while its accuracy is lower on InterCap, which presents the greatest challenges due to its complex human poses and strong human-object occlusions. In comparison to NICP, SegFit shows notably improved performance on EgoBody, where severe occlusions – such as instances where only a foot remains visible while the entire leg is hidden – highlight the importance of body-part segmentation as a critical signal for more accurate human registration. Lastly, while ArtEq is considerably faster due to its optimization-free nature, SegFit still offers a significant efficiency advantage over NICP, delivering at least a tenfold reduction in runtime while maintaining strong accuracy.

| Method | Metric | BEHAVE | EgoBody | Intercap |
|---|---|---|---|---|
| Human3D | Acc | 74.61% | 77.06% | 55.85% |
| + SegFit | | **91.22%** | **84.71%** | **67.86%** |
| Human3D | IoU | 57.38% | 62.89% | 45.37% |
| + SegFit | | **77.26%** | **69.48%** | **55.13%** |
| Human3D | mAP | 73.42% | 73.98% | 59.54% |
| + SegFit | | **85.74%** | **77.20%** | **67.69%** |

Table 2. Scores of human body part segmentation before and after SegFit. Metrics are accuracy (Acc), intersection over union (IoU), and mean average precision (mAP).

### 4.2. Refined Body-Part Segmentation

Beyond human registration, we also assess SegFit's enhanced body-part segmentation and compare it to the initial segmentation from Human3D [29], as shown in Table 2. After fitting the SMPL-X model, we refine the body-part labels by reassigning them through majority voting among the nearest model vertices, leading to improved segmentation quality. Human3D is first pre-trained on synthetic data and then fine-tuned on Kinect RGB-D sensor data, the same sensor used across all datasets, resulting in a minimal generalization gap. The improved results obtained by SegFit show that model-based segmentation still leads to substantial improvements in segmentation accuracy, of approximately 17%, 8%, and 12% on the BEHAVE, EgoBody, and InterCap datasets, respectively. This demonstrates that our method effectively enhances body-part segmentation by leveraging the fitted models. Figure 4 provides an example where Human3D initially misclassified a suitcase as the right leg. After applying SegFit, it was correctly identified as part of the background, demonstrating the effectiveness of our approach in resolving segmentation errors.

### 4.3. Self-Supervised Fine-tuning

We demonstrate that the refined body-part segmentation can be leveraged to fine-tune part segmentation networks in a self-supervised manner within a given dataset. This approach reduces the need for costly manual annotations when adapting to new, unseen datasets.

To this end, we use the outputs of SegFit as pseudo-ground truth to fine-tune Human3D and evaluate this strategy on BEHAVE [5] and InterCap [15]. The results are reported in Table 3. Fine-tuning is performed on a subset of each dataset, selected randomly. To minimize the impact of incorrectly segmented point clouds used for finetuning, we automatically exclude the 20% of samples for which SegFit exhibited the highest final loss values. Table 4 shows how this improves the mean V2V error and segmentation accuracy of the pseudo ground truths.

| Human3D Model | Metric | BEHAVE | InterCap |
|---|---|---|---|
| w/o Fine-Tuning | Accuracy | 74.61% | 55.85% |
| w/ Fine-Tuning | | **89.96%** | **67.98%** |
| w/o Fine-Tuning | IoU | 57.38% | 45.37% |
| w/ Fine-Tuning | | **75.36%** | **56.71%** |
| w/o Fine-Tuning | AP | 73.42% | 59.54% |
| w/ Fine-Tuning | | **84.56%** | **68.18%** |

Table 3. Segmentation performance of Human3D [29] before and after fine-tuning in a self-supervised manner on the outputs of the proposed SegFit.

On both tested datasets, the fine-tuning significantly improves Human3D both in terms of vertex-to-vertex error and segmentation accuracy. This demonstrates that using the proposed method for finetuning part segmentation networks on new data is highly beneficial.

| Dataset ↓ | V2V Orig. [mm] | V2V Filt. [mm] | Seg. Acc. Orig. [%] | Seg. Acc. Filt. [%] |
|---|---|---|---|---|
| **BEHAVE** | 37.4 | **30.1** | 90.44 | **92.35** |
| **InterCap** | 142.0 | **127.3** | 68.35 | **74.33** |

Table 4. Mean vertex-to-vertex (V2V) error and segmentation accuracy of pseudo ground truths before and after removing the 20% of point clouds for which SegFit exhibited the highest final loss value.

### 4.4. Analysis Experiments

Next, we conduct an ablation study to evaluate the impact of key components in our method, along with a hyperparameter analysis to examine the effect of loss weights.

**Ablation Study.** To assess the contributions of individual components of our method, we conduct an ablation study and show the results of four different variants in Table 5, where ④ is the full SegFit model.

① A fundamental yet informative baseline for SegFit is an optimization process that does not incorporate body-part segmentation. This baseline serves to isolate and quantify the specific contribution of segmentation to the overall performance. Instead of leveraging body-part segments, the model is fitted using a more generic approach, where it is initialized with four different orientations to address potential symmetry ambiguities. When compared to the full method, this segmentation-free approach results in a notable decline in performance across all evaluated metrics and datasets. These findings highlight the crucial role of body-part segmentation in improving human pose and shape estimation, demonstrating its effectiveness in guiding the optimization process toward more accurate results.

② The second baseline adds body-part segmentation during optimization while omitting the centroid-based initialization step. This modification leads to a significant re-


Figure 5. **Qualitative Results.** Example outputs of our SegFit on the EgoBody [38] dataset. From left to right: the input single-view point cloud showing the full scene including multiple humans, clutter and background, registered human meshes by SegFit from the front and side perspective, the refined body-part segmentation by SegFit. See Section 4.5 for additional details.

| | B.P. | Cent. | Metric | BEHAVE | EgoBody | InterCap |
|---|---|---|---|---|---|---|
| ① | ✗ | ✗ | | 55.6 | 109.7 | 206.8 |
| ② | ✓ | ✗ | V2V | 39.9 | 48.5 | 147.8 |
| ③ | ✗ | ✓ | in mm | 97.3 | 105.2 | 154.7 |
| ④ | ✓ | ✓ | | **37.0** | **47.9** | **147.2** |
| ① | ✗ | ✗ | | 43.6 | 100.3 | 195.4 |
| ② | ✓ | ✗ | MPJPE | 42.6 | 42.7 | 141.3 |
| ③ | ✗ | ✓ | in mm | 80.8 | 87.9 | 148.5 |
| ④ | ✓ | ✓ | | **42.2** | **20.6** | **140.7** |
| ① | ✗ | ✗ | | 6.05 | 4.28 | 10.7 |
| ② | ✓ | ✗ | Time | 2.98 | 3.70 | 6.21 |
| ③ | ✗ | ✓ | in s | **0.40** | **0.40** | **0.66** |
| ④ | ✓ | ✓ | | 1.86 | 1.79 | 1.37 |

Table 5. Ablation study analysing the effect of body parts (B.P.) and Centroids (Cent.) on the BEHAVE [5], EgoBody [38], and InterCap [15] datasets. Metrics are vertex-to-vertex (V2V) distance, mean-per-joint-position-error (MPJPE) and runtime.

duction in errors compared to the first baseline, which does not utilize segmentation, and brings the accuracy closer to that of the full method across all datasets. This variant also demonstrates a substantial boost in computational efficiency, reducing the runtime by at least a factor of two across all datasets. However, we observed that a key limitation of this approach is its slower convergence, which primarily arises due to misalignment of body limbs. Specifically, points from the inner side of an arm (or leg) can sometimes be incorrectly matched to points on the outer side (or vice-versa), leading to incorrect correspondences that hinder optimization. These misalignments can prolong the fitting process and reduce overall robustness. This observation motivated our introduction of centroid-based initialization, which provides a more structured starting point for optimization, improving both alignment accuracy and convergence speed.

③ Finally, we evaluate the accuracy of the centroid initialization step in isolation, without any subsequent optimization. While this approach results in an error approximately three times higher than the full method, it maintains a stable accuracy across all four datasets. Notably, it offers a significant speed advantage, with fitting times reduced by a factor of two to seven, averaging around half a second. This balance between speed and precision makes it a viable alternative for real-time applications where efficiency is prioritized over fine-grained accuracy.

**Hyper-parameter Analysis.** Table 6 examines the impact of the weighting coefficients $\lambda_{\text{pose}}$ and $\lambda_{\text{shape}}$ on the pose and shape terms. We conduct a grid search over the values $0.0, 0.5, 1.0, 2.0$ and find that the lowest V2V error is achieved when both coefficients are set to $0.5$. Notably, setting $\lambda_{\text{shape}} = 0$ underscores the importance of the shape term, which prevents unnatural body deformations.

| $\lambda_{\text{shape}} \backslash \lambda_{\text{pose}}$ | 0.0 | **0.5** | 1.0 | 2.0 |
|---|---|---|---|---|
| 0.0 | 52.2 | 45.0 | 47.6 | 48.7 |
| **0.5** | 51.6 | **41.8** | 42.5 | 44.1 |
| 1.0 | 51.6 | 42.9 | 44.3 | 45.2 |
| 2.0 | 51.8 | 44.1 | 45.7 | 46.9 |

Table 6. Effect of $\lambda_{\text{shape}}$ and $\lambda_{\text{pose}}$ on the vertex-to-vertex (V2V) error (in mm). Best value for both hyper-parameters is $0.5$.

### 4.5. Qualitative Results and Discussion

Figure 5 presents several representative examples of SegFit applied to the EgoBody [38] dataset. The input scenes and corresponding human poses exhibit significant diversity, introducing multiple challenges such as occlusions caused by scene clutter, partial visibility due to single-view depth sensors, and artifacts from the scanning process. In many cases, even for a human observer, it is difficult to discern which points belong to a person or to specific body parts based solely on the raw input data. Despite these challenges, our method demonstrates robust performance, successfully recovering human poses even under severe occlusions. The bottom example highlights this capability, emphasizing the importance of leveraging body-part segmentation to handle partial and noisy real-world point clouds. However, our approach is not without limitations. One common failure mode occurs when human instances are entirely missed during segmentation, resulting in missing reconstructions. Another frequent issue is inaccurate limb registration, particularly in cases where subjects cross their arms or legs, as seen in the second example from the top. These ambiguities can lead to incorrect alignments, particularly in highly occluded settings. To further enhance pose estimation in challenging scenarios, integrating additional scene reasoning, particularly with affordance-based constraints [11], could help refine predictions when subjects closely interact with their environment, where occlusions are most severe.

## 5. Conclusion

We introduce SegFit, a novel hybrid approach for fitting parametric human body models to diverse 3D point clouds, combining body part segmentation and human pose and shape priors to iteratively enhance both segmentation and pose fitting accuracy. Future work will explore potential improvements to the pose fitting accuracy, such as by introducing a penetration loss term for scenes where humans interact with each other or with objects in their environment. The code will be made publicly available.

## References


[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[2] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A Survey of Robot Learning from Demonstration. *Robotics and autonomous systems*, 2009. 1

[3] 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)*, 2020. 2, 5

[4] 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 *Neural Information Processing Systems (NeurIPS)*, 2020. 5

[5] 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 *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4, 5, 6, 8

[6] Aude Billard and Danica Kragic. Trends and Challenges in Robot Manipulation. *Science*, 2019. 1, 2

[7] 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 (ECCV)*, 2016. 1, 2

[8] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael Black. Dynamic faust: Registering human bodies in motion. pages 5573–5582, 2017. 4

[9] Zhongang Cai, Liang Pan, Chen Wei, Wanqi Yin, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. PointHPS: Cascaded 3D Human Pose and Shape Estimation from Point Clouds. *arXiv preprint arXiv:2308.14492*, 2023. 2

[10] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3D Human Pose from Structure and Motion. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[11] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. SceneFun3D: Fine-grained Functionality and Affordance Understanding in 3D Scenes. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[12] Haiwen Feng, Peter Kulits, Shichen Liu, Michael J Black, and Victoria Fernandez Abrevaya. Generalizing Neural Human Fitting to Unseen Poses with Articulated SE(3) Equivariance. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 4, 5

[13] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3D Hand Shape and Pose Estimation from a Single RGB Image. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[14] 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 (ICCV)*, 2019. 2, 3, 4

[15] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. *International Journal on Computer Vision (IJCV)*, 2024. 2, 4, 5, 6, 8

[16] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-End Recovery of Human Shape and Pose. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video Inference for Human Body Pose and Shape Estimation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop, 2019. 1, 2

[19] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep Iterative Matching for 6D Pose Estimation. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015. 1, 5

[21] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4

[22] Riccardo Marin, Enric Corona, and Gerard Pons-Moll. NICP: Neural ICP for 3D Human Registration at Scale. In *European Conference on Computer Vision (ECCV)*, 2025. 1, 2, 5

[23] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions On Graphics (TOG)*, 2017. 2

[24] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera Distance-aware Top-down Approach for 3D Multi-person Pose Estimation from a Single RGB Image. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[25] 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 *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 4

[26] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10318–10327, 2021. 2

[27] Edoardo Remelli, Artem Lukoianov, Stephan Richter, Benoit Guillard, Timur Bagautdinov, Pierre Baque, and Pascal Fua. MeshSDF: Differentiable Iso-surface Extraction. *International Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2

[28] Mel Slater and Maria V Sanchez-Vives. Enhancing our Lives With Immersive Virtual Reality. *Frontiers in Robotics and AI*, 2016. 1, 2

[29] Ayça Takmaz, Jonas Schult, Irem Kaftan, Mertcan Akçay, Bastian Leibe, Robert Sumner, Francis Engelmann, and Siyu Tang. 3D Segmentation of Humans in Point Clouds with Synthetic Data. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 5, 6

[30] Daniel Thalmann and Soraia Raupp Musse. *Crowd Simulation*. Springer Science & Business Media, 2012. 1, 2

[31] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep Autoencoder for Combined Human Pose Estimation and Body Model Upscaling. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[32] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning From Synthetic Humans. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[33] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally Aware Piecewise Transformation Fields for 3D Human Mesh Registration. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5

[34] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[35] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3S Human Shape and Articulated Pose Models. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[36] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4D Instance Segmentation of Close Human Interaction. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[37] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes-the Importance of Multiple Scene Constraints. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[38] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 4, 5, 7, 8

[39] Xinxin Zuo, Sen Wang, Qiang Sun, Minglun Gong, and Li Cheng. Self-Supervised 3D Human Mesh Recovery from Noisy Point Clouds. In *arXiv preprint arXiv:2107.07539*, 2021. 2