Locally Aware Piecewise Transformation Fields
for 3D Human Mesh Registration

Shaofei Wang[1], Andreas Geiger[2,3], Siyu Tang[1]

[1]ETH Zürich   [2]Max Planck Institute for Intelligent Systems, Tübingen   [3]University of Tübingen
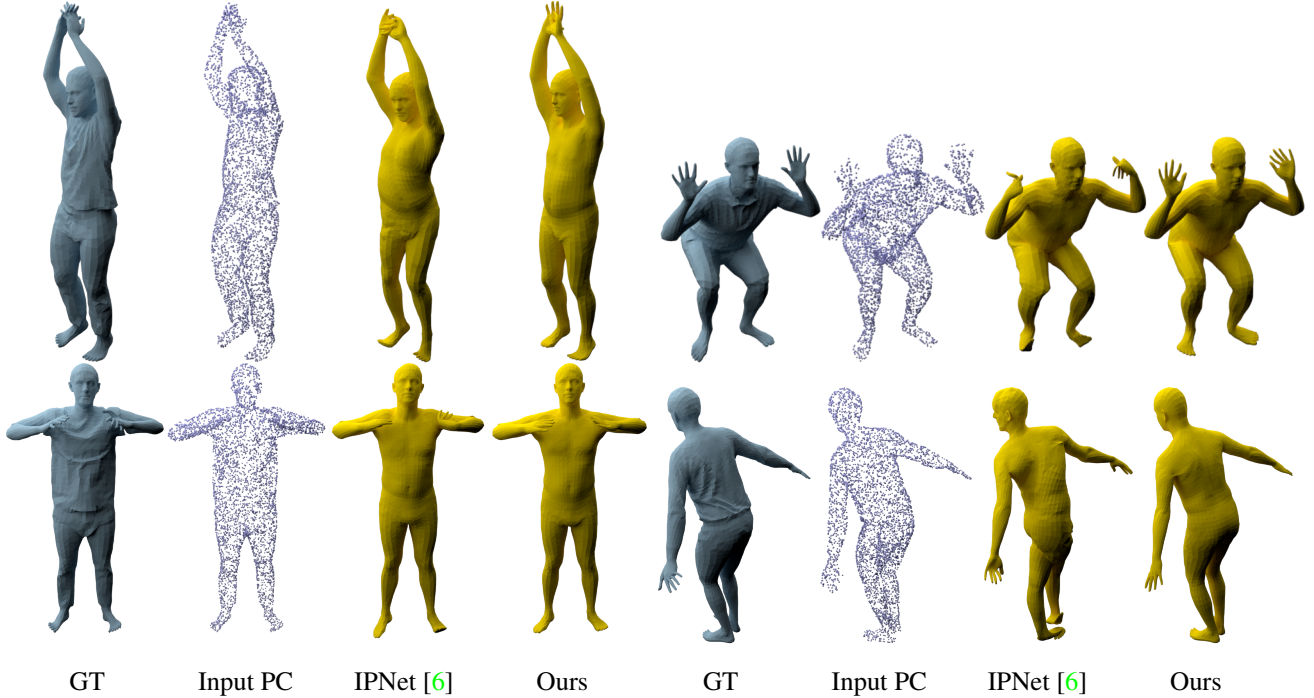
Figure 1: Our proposed approach aims at providing accurate pose initialization which facilitates the subsequent optimization-based registration. We show here ground-truth clothed meshes [30], input point clouds, under-cloth SMPL meshes registered by IPNet [6], and under-cloth SMPL meshes registered with our model.

Abstract

Registering point clouds of dressed humans to parametric human models is a challenging task in computer vision. Traditional approaches often rely on heavily engineered pipelines that require accurate manual initialization of human poses and tedious post-processing. More recently, learning-based methods are proposed in hope to automate this process. We observe that pose initialization is key to accurate registration but existing methods often fail to provide accurate pose initialization. One major obstacle is that, despite recent effort on rotation representation learning in neural networks, regressing joint rotations from point clouds or images of humans is still very chal-lenging. To this end, we propose novel piecewise trans-formation fields (PTF), a set of functions that learn 3D translation vectors to map any query point in posed space to its correspond position in rest-pose space. We com-bine PTF with multi-class occupancy networks, obtaining a novel learning-based framework that learns to simultane-ously predict shape and per-point correspondences between the posed space and the canonical space for clothed hu-man. Our key insight is that the translation vector for each query point can be effectively estimated using the point-aligned local features; consequently, rigid per bone trans-formations and joint rotations can be obtained efficiently via a least-square fitting given the estimated point corre-spondences, circumventing the challenging task of directly

regressing joint rotations from neural networks. Furthermore, the proposed PTF facilitate canonicalized occupancy estimation, which greatly improves generalization capability and results in more accurate surface reconstruction with only half of the parameters compared with the state-of-the-art. Both qualitative and quantitative studies show that fitting parametric models with poses initialized by our network results in much better registration quality, especially for extreme poses.

1. Introduction

Human pose and shape registration from sensor inputs is a long-standing problem in computer vision. While unstructured 3D point clouds are becoming increasingly available, accurately registering these point clouds to parametric human shape models [4, 18, 21, 28, 37, 57] still remains challenging, especially when considering clothed human with arbitrary poses. Traditional approaches often assume access to either temporal sequences of dense scans [10, 40, 59] or manually enforced constraints [3, 4, 9, 10, 18, 39, 41]. These requirements limit the applicability of traditional approaches to static, arbitrarily obtained point clouds.

With the advent of neural implicit functions [12, 31, 32, 36], learning-based methods that reconstruct human shapes from point clouds are becoming increasingly accurate [13]. However, most existing neural implicit models treat reconstructed human shapes as static objects and do not provide a way to register such reconstructions to parametric body models. Note that in the context of this paper, we refer to reconstruction as implicit surface reconstruction from point clouds, while registration refers to finding the shape and poses of a parametric model which best explains the input point cloud or the reconstructed surface. Some works distinguish registration (registering a point cloud to a template) [3, 4, 9, 17, 18, 55] and model-fitting (estimating parameters of a parametric model) [6, 7, 62]. In this context, our approach falls into the model-fitting category. However the definitions of registration and model-fitting are not consistent across papers, and for the ease of understanding we follow the convention of our major baseline [6] and use registration and model-fitting interchangeably.

More recently, IPNet [6] has been proposed for automatic point cloud registration of clothed humans. It predicts two sets of implicit surfaces, one for the clothed human and one for the human body under-cloth. IPNet then exploits optimization-based registration to fit a parametric model to the implicit surfaces from a fixed initial pose for all the subjects, with additional information about body part labels for each point in space. However, such semantic information is still very coarse and registration may fail when the target pose deviates too much from the initial pose (Fig. 1).

We observe that, although the local point cloud fea-

tures [13] lead to reliable surface reconstruction, which facilitates the registration of parametric human models (e.g. SMPL [28]), an accurate pose initialization is still key to reliable registration, as the underlying optimization problem is non-convex, thus reliable estimates are only obtained when initialized close to the solution. However, it is difficult to estimate parametric poses from point clouds directly, because regressing pose parameters from neural networks is hard and unintuitive (as investigated in [6], as well as demonstrated in our experiments in Sec. 5.3).

In this paper, we introduce a novel approach to estimating the SMPL [28] pose parameters, based on implicit representations and local point cloud features. Instead of regressing pose parameters directly from global features, we introduce a set of transformation functions, which take a query point and a local point cloud feature as input, and transform the query point to the rest-pose space (Fig. 2). We assign one such transformation function per body-part, and name these functions *Piecewise Transformation Fields* (PTF). The general idea of PTF is inspired by the observation that the rigid bone-transformations in SMPL can be calculated if we know point correspondences between the posed space and the unposed space. Another motivation for PTF is that by transforming query points into a canonical space before the occupancy query, we make the occupancy learning/inference task easier [14, 19]. With PTF, our novel occupancy networks can estimate for each point: 1) its double-layer occupancy value (*i.e.* inside body, in-between body and cloth, outside cloth) like IPNet does, 2) which body-part it belongs to, and 3) its corresponding position in rest-pose space. With this information, we can then extract a mesh-surface of human body, along with the semantic part label and the corresponding position in the rest-pose for each surface vertex. As a side-benefit, rigid transformations for each body-part can be estimated directly via least-square fitting. In terms of registering implicit surfaces to parametric models, our approach also employs optimization-based registration as IPNet. But unlike IPNet, our approach exploits point correspondences which allow us to more accurately initialize the pose parameters. This makes the registration process much more stable and accurate, especially for extreme poses; this will be evidenced in Section 5.

In summary, the contribution of this work is three-fold: (1) We propose Piecewise Transformation Fields (PTF) that learn to transform arbitrary points from posed space to rest-pose space. (2) We combine our PTF modules with occupancy networks, and achieve state-of-the-art results on clothed human reconstruction from point clouds on the CAPE dataset [30] while reducing the number of parameters by half. (3) We propose an alternative learning-based method for estimating joint rotations of the parametric SMPL model from point clouds. Our learning-based method takes advantage of local point-aligned fea-

tures and produces more accurate and robust estimations than direct regression from global features. When fitting parametric models to implicit reconstructions using our estimated poses as initialization, we achieve 18% reduction in registration error on average. Code is available at https://taconite.github.io/PTF/website/PTF.html.

The remainder of this paper is structured as follows: in Section 2 we give an overview of existing works that are related to our approach. In Section 3 we review the fundamentals, SMPL [28] and NASA [14], upon which we build our proposed approach. In Section 4 we introduce our proposed PTF and its application to fast joint rotation estimation. In Section 5 we benchmark our PTF model, showing its advantages for registration and reconstruction. In Section 6 we conclude and discuss possible future works.

2. Related Works

Optimization-based Registration: Optimization-based methods are most commonly used in traditional registration pipelines [3, 4, 18, 39, 40]. Traditional approaches rely on use of physical markers [3, 39], manually selected landmark-correspondences [4, 18] or colored patterns on skin [9, 10, 41], and are thus not practical for automatic registration of arbitrary point clouds. Most recently proposed automatic registration algorithms aim at fitting a parametric model to sparse keypoints detected from images [8, 37, 50], they usually require the design of a complicated non-linear, non-convex loss function, optimize this loss wrt. the body parameters (body shapes, rotations, global translation, etc.), and are often prone to local optima. There are also approaches that conduct optimization-based registration using dynamic 3D inputs, *e.g.* [10, 40, 59], to name a few. Since we focus on static point cloud registration here, a complete review on this aspect is beyond the scope of this paper.

Learning-based Registration: Learning-based methods have been proposed to enable automatic registration of clothed humans from sparse point clouds [6] or images [1, 2, 27, 54]. Earlier works focus on one-shot estimation of parametric models from images [15, 22–26, 35, 44, 46, 56, 58] using feed-forward neural networks. However, estimates from these methods are often noisy and most of them do not handle clothing. Recently, learning-based methods have been used to initialize important variables for optimization-based registration of clothed humans from point clouds [6] or images [1, 2, 54]. Our approach also uses a learning-based method to initialize optimization-based registration from sparse point clouds. Compared to the most recent learning-based registration approach for point clouds [6], our approach differs critically in that we extract additional point correspondence information, which produces more accurate pose initialization and thus improves subsequent optimization as evidenced

by our experiments. Note that our approach directly estimates pose/joint-rotations as initialization, which is orthogonal to landmark-constrained optimization presented in [4, 18, 54]. In a concurrent work [7], the authors also propose to conduct $\mathbb{R}^3$ to $\mathbb{R}^3$ correspondence prediction using neural networks. Our approach differs from theirs in several aspects: (1) [7] predicts correspondence from posed-space to unposed-unshaped-space and leaves pose estimation entirely to non-linear optimization, while our approach predicts correspondence from posed-space to unposed-shaped-space, enabling us to directly estimate joint rotations via fast linear optimization. (2) [7] does not handle surface reconstruction. (3) [7] devises a self-supervised training scheme, which, with moderate modifications, should be applicable to our framework as well, however we will leave such improvements for future studies.

Implicit Representations: With the introduction of neural implicit functions [12, 31, 32, 36], a number of implicit function-based methods have been proposed to reconstruct human bodies from either images [19, 47, 48] or point clouds [13]. However, all aforementioned methods treat human body as a rigid object and do not handle registration. [20, 34] model the human body as a non-rigid, deformable occupancy field and are able to transform human shape representations across poses via neural ordinary differential equations (ODEs) [11]. Our approach is partially inspired by [34] as both our approach and [34] learn a $\mathbb{R}^3$ to $\mathbb{R}^3$ transformation. However, our approach aims to transform points from posed space to rest-pose space and uses a traditional neural network to do that, while [20, 34] learn transformations from one posed-space to another posed-space. Furthermore, our approach utilizes local point cloud features to learn the transformation while [20, 34] are based on global features.

Correspondence Prediction for Model Fitting: Dense correspondence prediction has been employed for human pose estimation [42, 43, 53], hand pose tracking [52], non-rigid object alignment [5, 61] and undressed human mesh registration [17]. [42, 43, 53] regresses 3D correspondences from the input to a single canonical model after which they infer the human pose via non-linear optimization. [61] focuses on non-rigid object reconstruction and tracking and is not able to produce controllable models. [5] optimizes over a sparse set of points for general non-rigid object alignment but does not handle model-fitting of parametric models. [17] focuses on the registration of undressed people and does not handle dressed people registration or under-cloth body registration. Contrary to all these works, our approach addresses registrations of dressed humans and their under-cloth bodies, as well as SMPL parameter estimation simultaneously using a single neural-network.

3. Fundamentals

We start by briefly reviewing the fundamentals upon which we derive our approach: the classical SMPL model [28] and a recently proposed neural implicit approximation of the SMPL model, NASA [14].

3.1. SMPL Body Model

SMPL is a classical parametric human body model. In SMPL, shapes are controlled by a low-dimensional vector $\boldsymbol{\beta} \in \mathbb{R}^{10}$ which weights a set of different vertex-offsets that correspond to the principal components of PCA on training scans; poses for 24 joints (23 body-joints plus pelvis as root) are denoted as $\boldsymbol{\theta} \in \mathbb{R}^{72}$, where the pose for each joint is defined as the rotation (in axis-angle format) relative to its parent in the kinematic tree. To build a posed and shaped mesh of the human body, SMPL starts with an artist created mean-shape template $\bar{\mathbf{T}} \in \mathbb{R}^{3 \times 6890}$ with blend skinning weights $\mathcal{W} \in \mathbb{R}^{6890 \times 24}$. SMPL first applies a blend-shape function $B_S(\boldsymbol{\beta}) : \mathbb{R}^{|\boldsymbol{\beta}|} \mapsto \mathbb{R}^{3 \times 6890}$ which maps $\boldsymbol{\beta}$ to a set of additive per-vertex displacements that describes the person's body shape (shape-blend-shapes). Then, SMPL applies a pose-dependent blend shape function $B_P(\boldsymbol{\theta}) : \mathbb{R}^{|\boldsymbol{\theta}|} \mapsto \mathbb{R}^{3 \times 6890}$ which maps $\boldsymbol{\theta}$ to another set of additive per-vertex displacements that accounts for dynamic soft-tissue deformations caused by the pose deviation from the rest-pose (pose-blend-shapes). Adding the aforementioned displacements to the mean template results in a rest-pose mesh $\mathbf{T}_P(\boldsymbol{\beta}, \boldsymbol{\theta}) = \bar{\mathbf{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta})$, which includes the body shape of a person as well as dynamic soft-tissue deformations. Simultaneously, SMPL also regresses joint locations at the rest-pose from shape parameters via a linear joint regressor $J(\boldsymbol{\beta}) : \mathbb{R}^{|\boldsymbol{\beta}|} \mapsto \mathbb{R}^{72}$. A set of 4×4 rigid bone transformations $\{\mathbf{B}_b\} = \{\mathbf{B}_1, \cdots, \mathbf{B}_{24}\}$ can be constructed from rest-pose joint locations and relative joint rotations $\boldsymbol{\theta}$. SMPL applies a standard linear-blend-skinning (LBS) function $W(\cdot)$ with bone transformations $\{\mathbf{B}_b\}$ and skinning weights $\mathcal{W}$, transforming $\mathbf{T}_P(\boldsymbol{\beta}, \boldsymbol{\theta})$ into the posed and shaped mesh. Finally, global translation $\mathbf{t} \in \mathbb{R}^3$ is added to the posed and shaped mesh. In summary, SMPL models a function $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{t})$ that takes $\boldsymbol{\beta}, \boldsymbol{\theta}$ and $\mathbf{t}$ as inputs and outputs a human mesh:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{t}) = W(\bar{\mathbf{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}); \mathcal{W}, \{\mathbf{B}_b\}) + \mathbf{t}$$
(1)

We denote the 3×3 rotation matrix representation of joint b's rotation as $\mathbf{R}_b(\boldsymbol{\theta})$, with $\mathbf{R}_b(\mathbf{0}) = \mathbf{I}, \forall b \in \{1, \cdots, 24\}$. Let $\mathbf{t}_b \in \mathbb{R}^3$ be the location of rest-pose joint b obtained via $\mathcal{J}(\boldsymbol{\beta})$. Let $\mathbf{A}(b)$ denote the ordered set of b and its ancestors on the kinematic tree. Let $\bar{b}$ denote the parent of joint b. The bone-transformations $\{\mathbf{B}_b\}$ are defined as:

$$\mathbf{B}_b = G_b(\boldsymbol{\beta}, \boldsymbol{\theta}) G_b(\boldsymbol{\beta}, \mathbf{0})^{-1}$$
(2)

$$G_b(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{b' \in \mathbf{A}(b)} \begin{bmatrix} \mathbf{R}_{b'}(\boldsymbol{\theta}) & \mathbf{t}_{b'} - \mathbf{t}_{\bar{b}'} \\ \mathbf{0} & 1 \end{bmatrix}$$
(3)

$$\{\mathbf{t}_b\} = \mathcal{J}(\boldsymbol{\beta})$$
(4)

Note that adding the global translation $\mathbf{t}$ to the translation parts of $\{\mathbf{B}_b\}$ does not change the results of $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{t})$. Thus in the remainder of the paper we re-define $\{\mathbf{B}_b\}$ by adding $\mathbf{t}$ to its translation parts.

Registration: The SMPL model in Eq. (1) is end-to-end differentiable wrt. to $\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{t}$, thus given a point cloud, it is straightforward to construct an energy function on the distances between the point cloud and the SMPL mesh and minimize it wrt. $\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{t}$. However this optimization is non-convex and non-linear and thus prone to local minima. Therefore, manual initialization of pose and manual postprocessing to fix erroneous fits are usually required.

IPNet [6] proposes to reconstruct an inner surface (for under-cloth body) and an outer surface from point cloud of dressed humans. For registration, it first fits a SMPL mesh $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{t})$ to the inner surface, then it fits a SMPL+D [1, 27] mesh $M'(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{t}, \mathbf{D})$ to the outer surface, where $\mathbf{D}$ represents per-vertex displacements that capture clothing deformations in rest-pose space. Formally, [6] aims at minimizing the following energy:

$$E(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t}, \mathbf{D}) = w_{\text{data}} E_{\text{data}} + w_{\text{part}} E_{\text{part}} + w_{\text{reg}} E_{\text{reg}}$$
(5)

in which E_{data} measures the distances between reconstructed surfaces and the SMPL/SMPL+D meshes. E_{part} enforce part correspondences. E_{reg} is the regularization term, which includes a Laplacian regularization term for clothed surface, a pose prior from [40] and a shape prior that encourages $\boldsymbol{\beta}$ to be small. The E_{part} term makes the optimization more stable and thus [6] leads to improved registrations. However, the part correspondences are very coarse and the registration may fail for extreme poses as illustrated in Fig. 1.

In the next subsection we will review NASA [14], a recent work that relates neural implicit functions to Eq. (1).

3.2. Neural Articulated Shape Approximation

NASA [14] proposes to learn an implicit version of Eq (1). Towards this goal, NASA learns a set of occupancy functions parameterized by ω, $\{\bar{\mathcal{O}}_\omega^b\} = \{\bar{\mathcal{O}}_\omega^1, \cdots, \bar{\mathcal{O}}_\omega^{24}\}$, producing one occupancy value for each bone, given a 3D query point $\mathbf{x} \in \mathbb{R}^3$ and $\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{t}$[1]. The key idea of

---

[1]With some abuse of notation, we also use $\mathbf{x}$ to denote query points in homogeneous representation in $\mathbb{R}^4$ when applicable.

NASA is to learn the occupancy functions in the local-coordinate space of each bone with shape-blend-shapes and pose-blend-shapes. However these local-coordinates are piecewise *discontinuous*. We extend NASA's idea by learning occupancy functions in the *continuous* rest-pose space as this facilitates the correspondence learning that will be presented in Section 4. We refer readers to the Supp. Mat. for an elaboration on the difference between our modified NASA the original NASA [14]. For our modified NASA, $\{\bar{\mathcal{O}}_\omega^b\}$ is an implicit approximation to the $\bar{\mathbf{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta})$ term in Eq. (1). The occupancy value of an arbitrary point $\mathbf{x}$ in posed-space can be queried as:

$$\mathcal{O}(\mathbf{x}|\{\mathbf{B}_b\}, \mathbf{t}_1) = \max_b\{\bar{\mathcal{O}}_\omega^b\big(\mathbf{B}_b^{-1}\mathbf{x}, \boldsymbol{\Pi}_\omega^b[\{\mathbf{B}_b^{-1}\mathbf{t}_1\}]\big)\} \quad (6)$$

Where $\mathbf{t}_1 \in \mathbb{R}^4$ is the root-joint location in homogeneous representation. $\{\mathbf{B}_b^{-1}\mathbf{t}_1\} = \{\mathbf{B}_1^{-1}\mathbf{t}_1, \cdots, \mathbf{B}_{24}^{-1}\mathbf{t}_1\} \in \mathbb{R}^{72}$ is a compact representation of pose. $\boldsymbol{\Pi}_\omega^b \in \mathbb{R}^{72\times4}$ is a learned matrix that applies a sub-space linear projection on $\{\mathbf{B}_b^{-1}\mathbf{t}_1\}$, one per $b \in \{1, 2, \cdots, 24\}$. Since $\{\mathbf{B}_b\}$ depends on $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\mathbf{t}$ (Eq. (2)-(4)), the term $\boldsymbol{\Pi}_\omega^b[\{\mathbf{B}_b^{-1}\mathbf{t}_1\}]$ serves to provide shape-blend-shapes and pose-blend-shapes information to the occupancy functions $\{\bar{\mathcal{O}}_\omega^b\}$.

Essentially, Eq. (6) *canonicalizes* the query point $\mathbf{x}$ *before* doing the occupancy query. Thus the neural networks $\{\bar{\mathcal{O}}_\omega^b\}$ are learning occupancy in rest-pose, which reduces the difficulty for training occupancy functions compared to a posed representation. As we will show in Section 5, although NASA does not handle reconstruction, the general argument that canonicalizing query points helps occupancy learning still holds for our reconstruction networks.

4. Piecewise Transformation Fields

It is important to remark that NASA requires ground-truth SMPL paramaters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\mathbf{t}$ as inputs and does not solve the sensor-input-to-shape inference problem. In other words, NASA is a neural approximation to the SMPL model. In contrast, in this work, we assume $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\mathbf{t}$ and consequently $\{\mathbf{B}_b\}$ unknown and propose to learn a set of 3D point transformations from input space to canonical space. We parameterize these transformation functions by ψ, and formally define them as $\{\bar{\mathcal{T}}_\psi^b(\mathbf{x}, \mathbf{c_x})\}$: $(\mathbb{R}^3, \mathbb{R}^{|\mathbf{c_x}|}) \rightarrow \mathbb{R}^{3\times24}$. These functions approximate the *effect of bone transformations* on the query points instead of bone transformations themselves, *i.e.* $\bar{\mathcal{T}}_\psi^b(\mathbf{x}, \mathbf{c_x}) \approx \mathbf{B}_b^{-1}\mathbf{x}, \forall b \in \{1, 2, \cdots, 24\}$, where $\mathbf{c_x}$ is a point-aligned feature taken from a convolutional feature map [13, 38] at the location of $\mathbf{x}$. This is advantageous in that bone transformations are local for each bone, thus local features should be more suitable for predicting local bone transformations. This will be experimentally demonstrated in Section 5.3. Since we have one transformation field per bone, we formally name the set of transformation functions $\{\bar{\mathcal{T}}_\psi^b(\cdot, \cdot)\}$ *Piecewise Trans-*

formation Fields (PTF). With our PTF, the occupancy functions are now only dependent on point cloud features $\mathbf{c_x}$ and thus we can handle reconstruction from point cloud. Also, with the learned feature $\mathbf{c_x}$, we found that our networks generalize well even without the projection matrix $\boldsymbol{\Pi}_\omega^b$, thus we discard $\boldsymbol{\Pi}_\omega^b$ in our models.

With PTF, we can rewrite Eq. (6) as follows:

$$\mathcal{O}(\mathbf{x}|\mathbf{c_x}) = \max_b\{\bar{\mathcal{O}}_\omega^b\big(\bar{\mathcal{T}}_\psi^b(\mathbf{x}, \mathbf{c_x}), \mathbf{c_x}\big)\} \quad (7)$$

Note that this model differs critically from IPNet [6] in that our approach canonicalizes query points via PTF and performs occupancy queries in rest-pose space, while IP-Net does occupancy queries in posed-space. We denote the model defined in Eq. (7) as PTF-Piecewise as it is fully piecewise, *i.e.* the individual modules are independent of each other. We use fewer channels (80) per function as compared to IPNet (256), which, combined with the piecewise design, reduces the total number of parameters from 2.2M to 1.2M. As we will show in Section 5, although using fewer parameters, our PTF-Piecewise model retains a comparable reconstruction accuracy.

Loss functions: To define loss functions to Eq. (7), we first define the outputs from Eq. (7). We employ the multi-class occupancy outputs in IPNet which enable the occupancy networks to predict occupancy of dressed humans and their under-cloth bodies simultaneously, thus the set of outputs from $\{\bar{\mathcal{O}}_\omega^b\}$ can be denoted as $\{\hat{\mathbf{o}}_b\} \in [0, 1]^{3\times24}$. We further denote the outputs from $\{\bar{\mathcal{T}}_\psi^b\}$ as $\{\hat{\mathbf{x}}_b\} \in \mathbb{R}^{3\times24}$. Thus for each query point $\mathbf{x}$ we will predict 24 sets of occupancy value and correspondence. To obtain the final occupancy value $\mathcal{O}(\mathbf{x}|\mathbf{c_x}) = \hat{\mathbf{o}} \in [0, 1]^3$ and the correspondence $\mathcal{T}(\mathbf{x}|\mathbf{c_x}) = \hat{\mathbf{x}} \in \mathbb{R}^3$, we also need to define a probability simplex $\hat{\mathbf{p}} \in [0, 1]^{24}$, s.t. $\sum_{b=1}^{24} \hat{\mathbf{p}}_b = 1$ which describes part probabilities for the query point $\mathbf{x}$. This $\hat{\mathbf{p}}$ can be predicted via $\hat{\mathbf{p}}_b = \frac{\max \boldsymbol{\sigma}_b}{\sum_{b'} \max \boldsymbol{\sigma}_{b'}}$ where $\boldsymbol{\sigma}_b \in \mathbb{R}^3$ is the pre-activation occupancy logits of $\hat{\mathbf{o}}_b$. Note that we drop the dependencies of $\{\hat{\mathbf{o}}_b\}$, $\{\boldsymbol{\sigma}_b\}$, $\hat{\mathbf{o}}$ and $\hat{\mathbf{p}}$ on $\mathbf{x}$ for notation simplicity.

With the aforementioned notations, we can rewrite the outputs of Eq. (7) as:

$$\hat{\mathbf{o}} = \sum_{b=1}^{24} \hat{\mathbf{p}}_b\hat{\mathbf{o}}_b, \quad \hat{\mathbf{x}} = \sum_{b=1}^{24} \hat{\mathbf{p}}_b\hat{\mathbf{x}}_b \quad (8)$$

Similar to IPNet, we can also add an additional part classifier to predict $\hat{\mathbf{p}}$ directly from $\mathbf{c_x}$. This results in slightly more parameters than PTF-Piecewise but also slightly more accurate reconstruction than PTF-Piecewise. To keep the memory usage roughly the same as IPNet, we reduces the number of channels in the occupancy classifier from 256 to 128, and use 128 channels as well for the PTF. This reduces the total number of parameters in the decoder from 2.2M to 1.4M. For a fair comparison, we follow IPNet and use 14
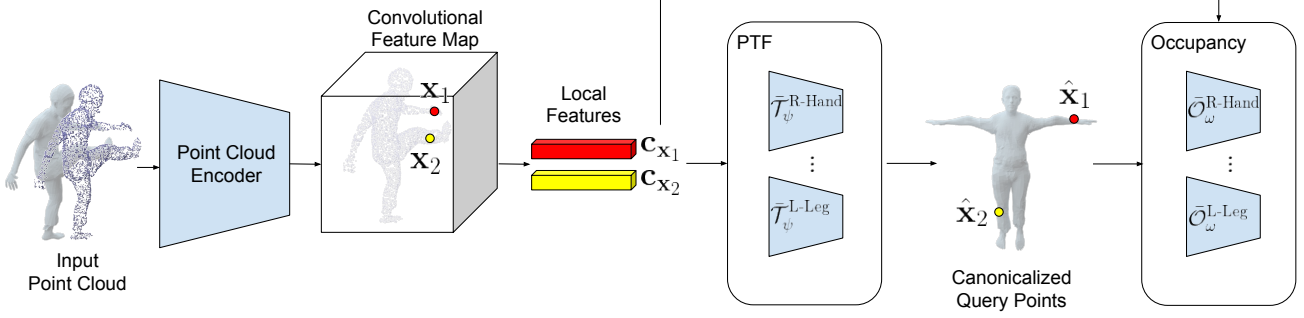
Figure 2: Overview of the proposed method. The key idea is to utilize the learned PTF module to transform query points $\mathbf{x}_1, \mathbf{x}_2$ in posed space into corresponding points $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2$ in unposed-space, and conduct occupancy classification in the unposed-space. This construction makes occupancy learning easier, while also providing dense $\mathbb{R}^3$ to $\mathbb{R}^3$ correspondences, which enables us to estimate joint-rotations via efficient linear optimization.

parts, *i.e.* we use 14 piecewise occupancy functions for the occupancy classifier. But the part classifier outputs 24 parts probabilities. This is necessary because SMPL defines 24 parts and in order to compute SMPL-compatible joint rotations, the labels for query points has to be consistent with SMPL. We merge the 24 parts probabilities into 14 parts in order to make Eq. (8) possible. We denote this model as the PTF-FC model. Please refer to the Supp. Mat. for detailed network architectures of PTF-Piecewise and PTF-FC.

To train our PTF models, we define three loss terms: an occupancy loss $\mathcal{L}_{occ}$, a classification loss $\mathcal{L}_{cls}$ and a correspondence loss $\mathcal{L}_{corr}$.

$$\mathcal{L}_{occ} = \text{CE}(\hat{\mathbf{o}}, \mathbf{o_x}) \qquad (9)$$

$$\mathcal{L}_{cls} = \text{CE}(\hat{\mathbf{p}}, \mathbf{p_x}) \qquad (10)$$

$$\mathcal{L}_{corr} = \left\| \hat{\mathbf{x}} - W^{-1}(\mathbf{x}; \mathcal{W}, \{\mathbf{B}_b\}) \right\|_2 \qquad (11)$$

where CE stands for multi-class cross-entropy loss. $\mathbf{o_x}$ is the ground-truth occupancy value for query point $\mathbf{x}$, similarly $\mathbf{p_x}$ is the one-hot vector of the ground-truth part-label. $W^{-1}(\mathbf{x}; \mathcal{W}, \{\mathbf{B}_b\}) : \mathbb{R}^3 \mapsto \mathbb{R}^3$ is the inverse LBS function that maps an arbitrary point from the posed-space to the rest-pose using ground-truth bone-transformations $\{\mathbf{B}_b\}$ and skinning weights $\mathcal{W}$. To obtain the ground-truth labels $\mathbf{p_x}$ and for calculating $W^{-1}(\mathbf{x}; \mathcal{W}, \{\mathbf{B}_b\})$ we need to associate the query point $\mathbf{x}$ to one of the 6890 SMPL vertices; this is done by assigning $\mathbf{x}$ its the nearest vertex on the SMPL mesh. Finally, we weight $\mathcal{L}_{cls}$ and $\mathcal{L}_{corr}$ with w_1 and w_2, the total loss is then:

$$\mathcal{L} = \mathcal{L}_{occ} + w_1 \mathcal{L}_{cls} + w_2 \mathcal{L}_{corr} \qquad (12)$$

we set $w_1 = 0.1$ and $w_2 = 1.0$ throughout our experiments.

Model Fitting with PTF: An advantage of our model is that it allows for efficient and accurate prediction of joint rotations. Let $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(D)}\}$ be a set of D points that lie on a human-mesh surface, and let $\mathbf{l} \in \{1, 2, \cdots, 24\}^D$ be the vector of semantic labels for these

D points. We estimate bone transformation $\mathbf{B}_b^*$ for b by:

$$\mathbf{B}_b^* = \arg\min_{\mathbf{B}_b} \sum_{i=1}^{D} [\mathbf{l}(i) = b] \|\mathbf{x}^{(i)} - \mathbf{B}_b \bar{\mathcal{T}}_\psi^b(\mathbf{x}^{(i)}, \mathbf{c}_{\mathbf{x}^{(i)}})\|_2$$

(13)

$$\mathbf{l}(i) = \arg\max_{b} \{\bar{\mathcal{O}}_\omega^b \big(\bar{\mathcal{T}}_\psi^b(\mathbf{x}^{(i)}, \mathbf{c}_{\mathbf{x}^{(i)}}), \mathbf{c}_{\mathbf{x}^{(i)}} \big)\} \qquad (14)$$

where $[\cdot]$ is an indicator function. Eq. (13) can be solved efficiently for each part via least-square fitting [51]. Joint rotation estimations $\boldsymbol{\theta}^*$ can then be factorized from $\mathbf{B}_b^*$ via Eq. (2)-(4). To make the estimation more robust to outliers, in practice we use least-square fitting Eq. (13) within RANSAC [16]. With the estimated $\boldsymbol{\theta}^*$ as initialization, we then fit SMPL/SMPL+D meshes to our reconstructed body and cloth surfaces according to Eq. (5) using exactly the same weights and learning rate schedules as in IPNet [6].

5. Experiments

Since IPNet [6] is the only existing automatic point cloud registration work for clothed humans, we primarily compare our approach against IPNet in terms of both registration quality and reconstruction quality. Furthermore, we provide ablation studies on encoder choices and the necessity of using our PTF architecture. We also compare joint rotation estimation of our approach to a naive baseline that regresses rotation matrices from global point cloud features.

Datasets: Because the official model of IPNet was trained on private data which is not publicly available, we instead use the CAPE dataset [30] to train and test IPNet and our models for a fair comparison. The CAPE dataset consists of 148584 pairs of registered clothed/minimally-clothed meshes of 15 subjects with different genders, the registered meshes all have SMPL topology and thus are compatible with SMPL. We use 12 subjects for training and 3 subjects for testing. We also subsample the recordings by a factor of 5. The final training set consists of 26004 frames while the

Method	Outer Err.	Inner Err.
IPNet [6]	28.2 mm	28.3 mm
PTF-Piecewise	23.9 mm	23.6 mm
PTF-FC-w/o-pose	27.5 mm	27.7 mm
PTF-FC	**23.1 mm**	**23.1 mm**

Table 1: Registration error on the full CAPE validation set (3965 frames). PTF-FC-w/o-pose indicates fitting SMPL/SMPL+D to our PTF-FC outputs without our pose initialization.

validation set consists of 3965 frames. We generate input point clouds by sampling 5K points on the surfaces of the clohted meshes with Gaussian noise of zero mean and 1mm standard-deviation. The sampling strategy of query points is as follows: we sample 2K points uniformly in the point cloud's bounding box, 4K points near clothed-meshes' surface and 4K points near the minimally-clothed-meshes' surface.

Training Details: We train all IPNet models and our models using the Adam optimizer with a learning rate of $1e-4$. The batch size is set to 12. We run 200K iterations for training, using a recently proposed convolutional point cloud encoder [38]. The training takes roughly two days on a single NVIDIA-V100 GPU.

Evaluations: In the remainder of this section, we will evaluate our approach along three dimensions. In Sec. 5.1 we evaluate results on registrations from point clouds to SMPL/SMPL+D meshes. In Sec. 5.2 we show how decoders with our PTF modules improve reconstruction quality while using less parameters than the IPNet decoder does. Lastly, in Sec. 5.3 we demonstrate the superiority of our local-feature-based pose estimation framework, which reduces error by a factor of two, compared to a baseline that regresses poses directly from global point cloud features.

5.1. Evaluation of Registration[2]

We follow the registration routine described in [6] to register from point clouds to SMPL and SMPL+D models for under-cloth bodies (inner) and clothed bodies (outer), respectively. For our PTF models, θ is initialized by solving Eq. (13) and factorizing out the rotation part of the results. Registration error is measured as average per-vertex euclidean distance from registered SMPL/SMPL+D mesh to the ground-truth SMPL/SMPL+D mesh. Errors are formally reported in Table 1.

As we can see in Table 1, on the entire CAPE validation set, our PTF achieve *18%* improvements on registration error compared to [6].

---

[2]For more results, please see the Supp. Mat.

Encoder	VRAM	mIoU	Outer CD	Inner CD
IFNet [13]	1.75G	86.7%	0.0191	0.0196
ConvONet [38]	1.25G	87.8%	0.0156	0.0162
ConvONet [38]+aug	**1.25G**	**88.6%**	**0.0151**	**0.0157**

Table 2: We fix the decoder to be the IPNet baseline and ablate on different encoders. VRAM indicates per-sample GPU memory usage during training. ConvONet achieves higher reconstruction quality with less memory usage.

5.2. Evaluation of Reconstruction

For evaluating reconstruction quality on the validation set, we report average intersection over union (mIoU $\uparrow$) of 100K random query points (50K sampled uniformly, 25K sampled near the clothed mesh, 25K sampled near the minimally-clothed mesh) and L1 Chamfer distance (CD) for both clothed surfaces (outer CD $\downarrow$) and under-cloth surfaces (inner CD $\downarrow$).

5.2.1 Ablation on Encoders

IPNet uses IFNet [13] as encoder that takes a voxel grid of size 128^3 as input, which we find very memory hungry and slow to run. With the aforementioned dataset setups, we can only use a batch size of 4. Given the number of iterations required, it would take weeks to train IPNet with 128^3 voxel inputs on the CAPE dataset. Reducing the voxel input to 64^3 helps with the training time, but it still takes about a week to train on CAPE. Instead, we choose to use a recently proposed convolutional feature encoder [38] with 3-plane-features, which we denote as ConvONet, as a more efficient and more accurate alternative to IFNet 64^3. A brief comparison of IFNet and ConvONet can be found in Table 2. Clearly, ConvONet is more practical on the CAPE dataset. We further apply data-augmentation that rotates the input point clouds randomly, which gives another boost in accuracy (ConvONet+aug in Table 2). We will thus use ConvONet as encoder and apply the random rotation to inputs during training throughout all remaining experiments.

5.2.2 Ablation on Decoders

We report reconstruction accuracy of IPNet and our PTF models in Table 3. PTF-Piecewise achieves comparable accuracy to IPNet while using only 54% parameters of IPNet. PTF-FC uses 64% parameters of IPNet, but achieves a noticeable improvement in clothed surface reconstruction.

We also ablate on different strategies of predicting point correspondences to demonstrate the necessity of using our PTF module. We first construct IPNet-Corr, which extends the final layer of IPNet's occupancy classifier to also predict the point correspondences $\hat{x}$ directly. This variant applies
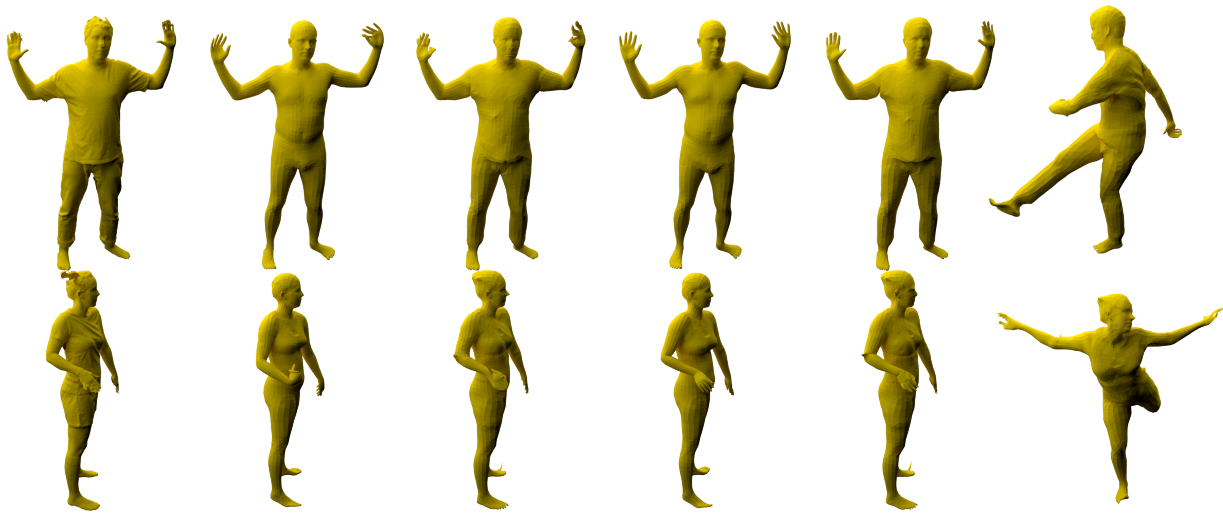
(a) Raw scan   (b) IPNet SMPL   (c) IPNet SMPL+D   (d) Ours SMPL   (e) Ours SMPL+D   (f) Ours re-posed

Figure 3: Although our model is trained on point clouds sampled from registered meshes, it generalizes well to raw scans. We show qualitative results on the BUFF dataset [59]. Please see the Supp. Mat. for more results.

Method	#params	mIoU	Outer CD	Inner CD
IPNet [6]	2.2M	88.6%	0.0151	**0.0157**
PTF-Piecewise	**1.2M**	89.4%	0.0152	0.0162
PTF-FC	1.4M	**89.7%**	**0.0148**	0.0158
Ablation on Point Correspondence Prediction				
Method	#params	mIoU	Outer CD	Inner CD
IPNet+Corr	2.3M	88.9%	0.0153	0.0159
TF-FC	7.3M	88.1%	0.0154	0.0161

Table 3: Ablation study on decoder architectures. Our PTF models achieve comparable or better reconstruction quality while using less parameters than the baseline.

Method	Per-vetex error
Baseline	74.4 mm
PTF-Piecewise	34.8 mm
PTF-FC	**34.1 mm**

Table 4: Comparison of our pose estimation approach against the baseline that regresses a continuous 6D representation [60] of rotation matrices directly from global point cloud features.

occupancy inference in the posed-space. IPNet+Corr gives slightly better mIoU but worse Chamfer distance than the IPNet baseline.

Next, we construct a variant, TF-FC. TF-FC replaces the PTF module with a 4-layer MLP with 128×14 channels. This simplification results in degraded performance.

5.3. Evaluation of Rotation Estimation

In this experiment, we demonstrate superiority of our local-feature-based pose estimation pipeline versus a direct regression baseline. The baseline takes 5K points as input, encodes the point cloud with PointNet [45], and then regresses rotation matrices in 6D continuous representation [60]. We evaluate the per-vertex error on the SMPL mesh, assuming ground-truth shape and global-translation are known. For our method, θ^* is obtained by solving Eq. (13)-(14) without optimizing Eq. (5). Results are shown in Table 4. We reduce the per-vertex error significantly, from 74.4mm to 34.1mm.

6. Conclusion

In this paper, we introduce Piecewise Transformation Fields (PTFs), a set of piecewise transformation fields that learns to canonicalize query points. Combining PTFs and occupancy networks results in more parameter-efficient models that not only improve point cloud reconstruction quality, but also enable us to estimate joint rotations accurately through efficient optimization. With better reconstructed surfaces and accurately initialized poses, we achieve state-of-the-art results on automatic point cloud registration of dressed people. In the future, we plan to extend the proposed PTFs to the image domain, as well as combine it with other neural-representations for human bodies [29, 33, 49] that require poses as inputs.

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 *Proc. of CVPR*, 2019. 3, 4

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proc. of ICCV*, 2019. 3

[3] Brett Allen, Brian Lee Curless, and Zoran Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 22(3), 2003. 2, 3

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 24, 2005. 2, 3

[5] Florian Bernard, Zeeshan Khan Suri, and Christian Theobalt. Mina: Convex mixed-integer programming for non-rigid shape alignment. In *Proc. of CVPR*, 2020. 3

[6] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Proc. of ECCV*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[7] 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 *Proc. of NeurIPS*, 2020. 2, 3

[8] 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 *Proc. of ECCV*, 2016. 3

[9] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proc. of CVPR*, 2014. 2, 3

[10] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *Proc. of CVPR*, 2017. 2, 3

[11] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Proc. of NeurIPS*, 2018. 3

[12] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proc. of CVPR*, 2019. 2, 3

[13] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proc. of CVPR*, 2020. 2, 3, 5, 7

[14] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa: Neural articulated shape approximation. In *Proc. of ECCV*, 2020. 2, 3, 4, 5

[15] Endri Dibra, Himanshu Jain, Cengiz Oztireli, Remo Ziegler, and Markus Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *Proc. of CVPR*, 2017. 3

[16] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of ACM*, 24(6), 1981. 6

[17] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. 3d-coded : 3d correspondences by deep deformation. In *Proc. of ECCV*, 2018. 2, 3

[18] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A Statistical Model of Human Pose and Body Shape. *Computer Graphics Forum*, 28:337–346, 2009. 2, 3

[19] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proc. of CVPR*, 2020. 2, 3

[20] Chiyu Max Jiang, Jingwei Huang, Andrea Tagliasacchi, and Leonidas Guibas. Shapeflow: Learnable deformations among 3d shapes. In *Proc. of NeurIPS*, 2020. 3

[21] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proc. of CVPR*, 2018. 2

[22] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc.of CVPR*, 2018. 3

[23] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proc. of CVPR*, 2020.

[24] 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 *Proc. of ICCV*, 2019.

[25] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proc. of CVPR*, 2019.

[26] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proc. of CVPR*, 2017. 3

[27] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *Proc. of 3DV*, 2019. 3, 4

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 34(6), 2015. 2, 3, 4

[29] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proc. of CVPR*, 2021. 8

[30] 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 *Proc. of CVPR*, 2020. 1, 2, 6

[31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. of CVPR*, 2019. 2, 3

[32] Mateusz Michalkiewicz, Jhony K. Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proc. of ICCV*, 2019. 2, 3

[33] Marko Mihajlovic, Yan Zhang, Michael J. Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proc. of CVPR*, 2021. 8

[34] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by

learning particle dynamics. In *Proc. of ICCV*, 2019. 3

[35] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *Proc. of 3DV*, 2018. 3

[36] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. of CVPR*, 2019. 2, 3

[37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proc. of CVPR*, 2019. 2, 3

[38] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proc. of ECCV*, 2020. 5, 7

[39] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 2017. 2, 3

[40] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. 2, 3, 4

[41] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 34(4), Aug. 2015. 2, 3

[42] Gerard Pons-Moll, Jonathan Taylor, Jamie Shotton, Aaron Hertzmann, and Andrew Fitzgibbon. Metric regression forests for human pose estimation. In *Proc. of BMVC*, 2013. 3

[43] Gerard Pons-Moll, Jonathan Taylor, Jamie Shotton, Aaron Hertzmann, and Andrew Fitzgibbon. Metric regression forests for human pose estimation. *International Journal of Computer Vision*, 113, 2015. 3

[44] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *Proc. of CVPR*, 2017. 3

[45] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. of CVPR*, 2017. 8

[46] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *Proc. of ICCV*, 2019. 3

[47] 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 *Proc. of ICCV*, 2019. 3

[48] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proc. of CVPR*, 2020. 3

[49] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proc. of CVPR*, 2021. 8

[50] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *Proc. of ECCV*, 2020. 3

[51] Olga Sorkine-Hornung and Michael Rabinovich. Least-squares rigid motion using svd, 2017. 6

[52] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, Arran Topalian, Erroll Wood, Sameh Khamis, Pushmeet Kohli, Toby Sharp, Shahram Izadi, Richard Banks, Andrew Fitzgibbon, and Jamie Shotton. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 35(4), 2016. 3

[53] Jonathan Taylor, Jamie Shotton, Toby Sharp, and Andrew Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Proc. of CVPR*, 2012. 3

[54] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proc. of ECCV*, 2018. 3

[55] Lingyu Wei, Qixing Huang, Duygu Ceylan, Etienne Vouga, and Hao Li. Dense human body correspondences using convolutional networks. In *Proc. of CVPR*, 2016. 2

[56] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proc. of CVPR*, 2019. 3

[57] 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 *Proc. of CVPR*, 2020. 2

[58] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proc. of ICCV*, 2019. 3

[59] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proc. of CVPR*, 2017. 2, 3, 8

[60] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *Proc. of CVPR*, 2019. 8

[61] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rhemann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, and Marc Stamminger. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 33(4), 2014. 3

[62] Silvia Zuffi and Michael J. Black. The stitched puppet: A graphical model of 3D human shape and pose. In *Proc. of CVPR*, 2015. 2