ETCH: Generalizing Body Fitting to Clothed Humans via Equivariant Tightness

Boqian Li[1] Haiwen Feng[2,3*] Zeyu Cai[1] Michael J. Black[2] Yuliang Xiu[1,2†]

[1]Westlake University [2]Max Planck Institute for Intelligent Systems [3]UC Berkeley

{liboqian, caizeyu, xiuyuliang}@westlake.edu.cn {haiwen.feng, black}@tuebingen.mpg.de

boqian-li.github.io/ETCH

Figure 1. **Body Fitting on Clothed Humans.** Given 3D clothed humans in any pose and clothing, ETCH accurately fits the *body underneath*. Our key novelty is modeling cloth-to-body SE(3)-equivariant tightness vectors for clothed humans, abbreviated as ETCH, which resembles "etching" from the outer clothing down to the inner body. The ground-truth body is shown in blue, our fitted body in green, and ground-truth markers as ●.

Abstract

Fitting a body to a 3D clothed human point cloud is a common yet challenging task. Traditional optimization-based approaches use multi-stage pipelines that are sensitive to pose initialization, while recent learning-based methods often struggle with generalization across diverse poses and garment types. We propose Equivariant Tightness Fitting for Clothed Humans, or ETCH, a novel pipeline that estimates cloth-to-body surface mapping through locally approximate SE(3) equivariance, encoding tightness as displacement vectors from the cloth surface to the underlying body. Following this mapping, pose-invariant body features regress sparse body markers, simplifying clothed human fitting into an inner-body marker fitting task. Extensive experiments on CAPE and 4D-Dress show that ETCH significantly outperforms state-of-the-art methods – both tightness-agnostic and tightness-aware – in body fitting accuracy on loose clothing ($16.7\% \sim 69.5\%$) and shape accuracy (average 49.9%). Our equivariant tightness design can even reduce directional errors by ($67.2\% \sim 89.8\%$) in one-shot (or out-of-distribution) settings ($\approx 1\%$ data). Qualitative results demonstrate strong generalization of ETCH, regardless of challenging poses, unseen shapes, loose clothing, and non-rigid dynamics. We will release the code and models soon for research purposes at boqian-li.github.io/ETCH.

1. Introduction

Fitting a template body mesh to a 3D clothed human point cloud is a long-standing challenge, which is vital for shape matching, motion capture, and animation, while enabling numerous applications like virtual try-on and immersive teleportation. It also supports the development of parametric human body models such as SMPL [37] and GHUM [64]. Aligning a template body mesh with raw human scans, *i.e.*, unordered point clouds, provides a common topology supporting statistical modeling of body shape and deformation.

Traditional optimization-based fitting pipelines (details in Sec. 2.1) are complex, depending on many ad-hoc rules to address corner cases, such as inconsistent 2D keypoints and additional skin-clothing separation. A path forward would leverage 3D scans paired with accurate underlying bodies to train a data-driven 3D pose regressor. However, despite advances in point-based neural networks, like PointNet [51] and PointNet++ [52], they still struggle with fitting tasks, particularly in achieving **robust generalization** under the complexity of *varied human poses and shapes, various clothing styles, and non-rigid dynamics*.

In recent works on fitting a parametric body model to

*Project Lead
†Corresponding Author

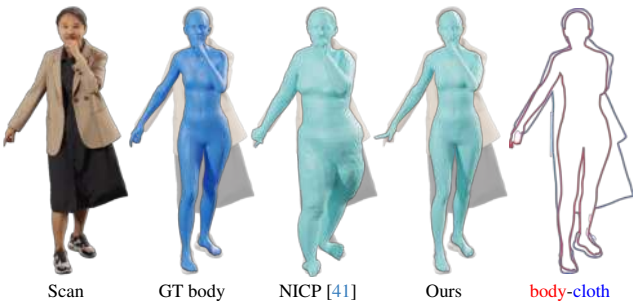

Figure 2. **Registration vs. Fitting.** Though both registration and fitting involve placing body inside clothing, "registration", like NICP [41], focuses on matching the *outer surface*, whereas "fitting" emphasizes aligning with the *underlying body*, making it more robust to clothing variations.

scans, ArtEq [22] addresses out-of-distribution pose generalization through articulated SE(3) equivariance. However, while ArtEq is designed to tackle articulation, it struggles with significant clothing deviations from the underlying body, particularly with loose garments like skirts or large deformations caused by dynamic motions, as shown in Fig. 1. Other recent approaches attempt to address this by separating garment layers from the underlying body, using scalar tightness [15] or double-layer occupancy [7, 58]. These methods simplify fitting loosely clothed humans by reducing the problem to a bare-body fitting one. However, given the combinatorial deformations caused by various human motions and clothing, these methods still struggle with out-of-distribution poses or shapes, as *Sup.Mat.*'s Fig. R.4 shows.

We integrate the two – "equivariance" and "tightness", capitalizing on their strengths to achieve the best of both worlds, and present *Equivariant Tightness Fitting for Clothed Humans (ETCH)*. Specifically, ETCH encodes tightness as displacement vectors from the observed cloth surface to the underlying body surface, capturing their dynamic interaction. This encoding remains approximately locally SE(3) equivariant under clothing deformation, accommodates both tight- and loose-fitting garments, and consistently points toward the inner body regardless of pose and clothing deformations. The encoding is learned from ground-truth tightness vectors computed from paired data, *i.e.*, the clothed human meshes paired with the underlying ground-truth body.

As shown in Fig. 2, during inference the predicted tightness vectors ($\mathbf{V}$) use the outer surface points to approximate the inner points that define the body's shape. However, this only accomplishes half the task of fitting the body, as the points remain unordered and do not offer a direct correspondence. To ensure robust body fitting, we perform sparse marker regression, which is more efficient and less sensitive to outliers compared to dense correspondence methods like PTF [58] (see Tab. 2). In this approach, we aggregate the dense inner points into a sparse set of virtual markers on the body surface. As shown in Fig. 3, each marker is assigned a region label ($\mathbf{L}$) corresponding to a specific area of the body, and the dense inner points sharing the same label—those clustered within the same region—are weighted and aggregated based on their predicted confidence ($\mathbf{C}$). This process yields a robust localization of the markers on the body surface. Consequently, fitting a parametric body to these markers is a well-solved problem [38].

And unlike direct body regression methods like ArtEq [22], our approach is tightness-aware. By correctly disentangling cloth deformation from body structure in an equivariant manner and regressing the sparse markers with invariant features, ETCH demonstrates strong generalization across challenging poses, extreme shapes, diverse garments, and clothing dynamics, supported by extensive qualitative and quantitative results. Each design choice has been validated through ablations.

Our key contributions are as follows:

- **Equivariant Tightness Fitting.** We introduce a novel framework, abbreviated as ETCH, which models cloth-to-body mapping using *tightness-equivariant displacement vectors* and leverages *pose-invariant body correspondences* and *weighted aggregation* for accurate sparse marker placement. This reformulates the clothed human fitting into a tightness-aware marker-based fitting task.
- **Superior Performance.** ETCH significantly outperforms prior approaches (*i.e.*, IPNet, PTF, NICP, ArtEq) across datasets and metrics (see Tab. 1). It shows superior *generalization* to loose dynamic clothing and challenging poses and extreme shapes at *Sup.Mat.*'s Figs. R.3 and R.4, and shows strong out-of-distribution generalization at Fig. 7 and Tab. 3.
- **In-depth Analysis.** We perform extensive ablation studies to assess each component's contribution, including tightness (direction vs. scalar) at Tab. 2, sparse markers vs. dense correspondences at Tab. 2 and Fig. 6, the role of equivariance features at Fig. 7 and Tab. 3, and the effectiveness of post-refinement at *Sup.Mat.*'s Tab. R.2.

2. Related Work

2.1. Body Fitting for 3D Clothed Humans

Estimating body shape under clothing has a long history. Balan and Black [5] introduce the problem of fitting a parametric 3D human body model *inside* a clothed 3D surface. They work with a coarse visual-hull representation extracted from multi-view images or video while later work uses 3D scans [26, 54, 62, 65]. All these methods rely on the same principle of optimizing for a single parametric body shape that, when posed, fits inside all the scans. None of them explicitly model how clothing deviates from the body. The first work to learn a mapping between the inner body shape and outer, clothed, shape does so in 2D [25] by building a statistical model of how clothing deviates from the body.

Surface Registration vs. Body Fitting. Modeling the cloth-to-body mapping relates to scan fitting terminology. "Surface

Registration" and "Body Fitting" are often used interchangeably but describe distinct processes. As illustrated in Fig. 2, for 3D clothed humans, body fitting aims to deform a body model (*e.g.*, SMPL [37]) to align with the *underlying body*, optimizing within the model space (*e.g.*, SMPL's pose and shape parameters). In contrast, registration (or "alignment") deforms a template mesh to match the *outer surface*, optimizing outside the model space (*e.g.*, via displacements from the body, like SMPL-D [7, 39, 48, 69]). These terms can be interchangeable when clothing is minimal [10, 27, 33]. However, for loose clothing, "tightness-agnostic" registration will inflate the human shapes to align with the outer clothing surface, as shown in Fig. 2 and *Sup.Mat.*'s Fig. R.4.

With ETCH, our goal is to achieve generalizable *body fitting* for 3D clothed humans by separating clothing from the underlying SMPL body shape. Our approach follows the "tightness-aware" paradigm (see Tab. 1) established by TightCap [15], IPNet [7], and PTF [58]. Prior work has exploited the fact that the body should fit tightly to the scan in skin regions but more loosely in clothed regions [44, 69]. Zhang et al. [69] use a binary tightness model in which the body "snaps" to nearby clothing surfaces but ignores distant ones. Neophytou and Hilton [43] introduce a learned 3D model of clothing displacement from a parametric body shape, but this model has to be trained for each actor. Our key innovation is defining tightness as a set of vectors, rather than relying on a scalar UV map [15] or double-layer occupancy [7]. Additionally, we conduct the fitting in the posed rather than canonical space like PTF [58], being more robust to raw 3D human captures with out-of-distribution poses.

Optimization vs. Learning. Prior work could also be grouped into "Optimization-based" and "Learning-based" approaches. Optimization approaches refine results iteratively using the ICP algorithm and its variants [1, 16, 47, 76], often aided by additional cues such as markers [1, 2, 46] or color patterns [10, 12]. However, their sensitivity to noise and poor initial poses can hinder convergence, resulting in suboptimal local minima or failed fitting.

In particular, the most commonly used optimization-based body fitting pipelines [6, 39, 44, 67, 69, 71] typically proceed as follows: multiview rendering of scans $\Rightarrow$ 2D keypoint estimation (*e.g.*, OpenPose [13], AlphaPose [21], *etc.*) $\Rightarrow$ triangulation of 3D keypoints from multiview 2D keypoints $\Rightarrow$ fitting SMPL (-X) parameters using 3D keypoints and the 3D point cloud (often accompanied by skin segmentation [3, 24] or self-intersection penalties [42]). 2D keypoint estimation may fail in loose clothing, and triangulation based on such inconsistent 2D keypoints can lead to skewed 3D joint estimates. This harms the fitting process, as chamfer-based optimization is highly sensitive to pose initialization and can easily get trapped in local minima.

Learning-based methods, powered by the large-scale 3D datasets (*e.g.*, AMASS [40]) and point cloud networks (*e.g.*, PointNet [51, 52], DeepSets [68], KPConv [55], and PointTransformer [60, 61, 70]), either improve initialization for subsequent optimization steps [7, 58], directly output meshes [49, 57, 73], or directly predict the pose/shape parameters of statistical body models [8, 29, 36, 73]. However, direct model-based regression is difficult [7, 8, 29, 57, 58]. To circumvent this, current approaches use intermediate representations, such as joint features [29, 36], correspondence maps [8, 58], segmentation parts [7, 8, 22, 22, 58], or resort to temporal consistency [69] or motion models [28].

2.2. SE(3)Equivariant and Human Pointcloud

Convolutional neural networks (CNNs) owe much of their success to translational equivariance. Building on this, researchers have explored SE(3)-equivariant neural networks for point cloud processing, including Vector Neurons [19], Tensor Field Networks [56], SE3-transformers [23], and methods leveraging group averaging theory or SO(3) discretization. For example, EPN [14] discretizes SO(3) and applies separable discrete convolutions (SPConv) for computational efficiency, simplifying rigid-body pose regression.

Extending to non-rigid scenarios, recent work addresses part-level articulation at the object or scene level (*e.g.*, [20, 50, 72]). Human bodies represent a special case due to their high articulation. Efforts like Articulated SE(3) Equivariance (ArtEq) [22] or APEN [4] model local or piecewise SO(3) equivariance to handle articulated deformations. However, while the human skeleton is articulated, loose clothing does not strictly follow these assumptions: garments can undergo large, complex deformations that violate standard rigidity or articulation models. Consequently, achieving robust body fitting via approximate SE(3) equivariance for loosely clothed humans remains an open challenge.

In contrast, we address this challenge from a new perspective by modeling approximate equivariance between the cloth surface and body surface through the local SO(3) equivariance of a "tightness" vector. This approach enables explicit cloth-to-body modeling while taking advantage of approximate equivariance to achieve more robust body fitting under loose clothing, effectively combining the strengths of learning-based pointcloud-to-marker regression and optimization-based marker-to-body fitting.

3. Method

Given a point cloud $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^3\}_N$, which is randomly sampled from the 3D humans, our goal is to optimize the pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$ and $\mathbf{t}$ parameters of SMPL [37] (Sec. 3.1), via a proxy task – fitting SMPL to the estimated sparse body markers, denoted as $\hat{\mathbf{M}} = \{\hat{\mathbf{m}}_k \in \mathbb{R}^3\}_K$, here $K = 86$. To accurately estimate these body markers, we leverage the SE(3)-equivariant & invariant pointwise features (Sec. 3.1). Using these pointwise features as input, we regress three key components to model the *cloth-to-body tightness*, re-

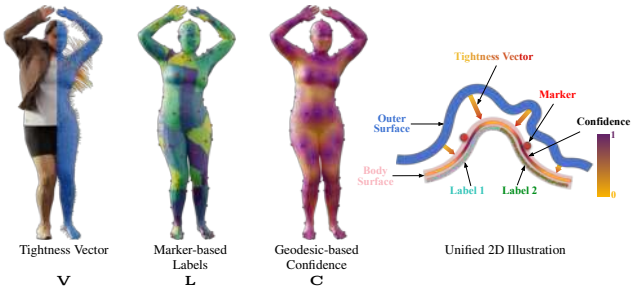

Figure 3. **Terminology of Tightness-Vector and Marker-Confidence.** We illustrate the key components used for data preparation: 1) Tightness Vectors $\mathbf{V}$, which connect the outer surface points with underneath body, and transmitting 2) Marker-based Labels $\mathbf{L}$ and Confidence $\mathbf{C}$. We also provide a 2D illustration that unifies these terms together. Sparse markers as ●, and confidence bar indicates the geodesic distance to the closest marker.

fer to Fig. 3 for illustrations: **1) a vector** $\mathbf{v}_i$ – pointing from the outer surface point cloud to the inner body surface (Sec. 3.3), constructed with the directional term $\mathbf{d}_i$ and magnitude $b_i = \|\mathbf{v}_i\|$, **2) a label** l_i – indicating the corresponding inner marker of $\mathbf{x}_i$, **3) a confidence** c_i – quantifies the contribution of inner points $\mathbf{y}_i$ in marker aggregation (Sec. 3.4). With these elements, we apply weighted (with estimated confidence) aggregation with tightness vector clusters of each marker label, to accurately regress the ultimate body markers $\hat{\mathbf{M}}$, which are finally used to optimize the final body model (Sec. 3.5). How to prepare the ground-truth components $\{\mathbf{V}, \mathbf{L}, \mathbf{C}\}$ for training is detailed in Sec. 3.2.

3.1. Preliminary

Parametric Body Model – SMPL. SMPL [37] has been a standard body format in various clothed human datasets [39, 59, 67, 71]. It is a statistical model that maps shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and pose $\boldsymbol{\theta} \in \mathbb{R}^{J\times 3}$ parameters to mesh vertices $\mathbf{V} \in \mathbb{R}^{6890\times 3}$, where J is the number of human joints ($J = 24$). $\boldsymbol{\beta}$ contains the blend weights of shape blendshapes B_S, and $B_S(\boldsymbol{\beta})$ accounts for variations of body shapes. $\boldsymbol{\theta}$ contains the relative rotation (axis-angle) of each joint plus the root one w.r.t. their parent in the kinematic tree, and $B_P(\boldsymbol{\theta})$ models the pose-dependent deformation. Both shape displacements $B_S(\boldsymbol{\beta})$ and pose correctives $B_P(\boldsymbol{\theta})$ are added together onto the template mesh $\bar{\mathbf{T}} \in \mathbb{R}^{6890\times 3}$, in the rest pose (or T-pose), to produce the output mesh $\mathbf{T}$:

$$\mathbf{T}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \bar{\mathbf{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}), \tag{1}$$

Next, the joint regressor $J(\boldsymbol{\beta})$ is applied to the rest-pose mesh $\mathbf{T}$ to obtain the 3D joints : $\mathbb{R}^{|\boldsymbol{\beta}|} \rightarrow \mathbb{R}^{J\times 3}$. Finally, Linear Blend Skinning (LBS) $W(\cdot)$ is used for reposing purposes, the skinning weights are denoted as $\mathcal{W}$, then the posed mesh is translated with $\boldsymbol{t} \in \mathbb{R}^3$ as final output $\mathbf{M}$:

$$\mathbf{M}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{t}) = W(\mathbf{T}(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \boldsymbol{t}, \mathcal{W}). \tag{2}$$

Local SE(3) Equivariance Features. A function $\mathcal{F}$ is equivariant if, for any transformation $\mathcal{T}$, $\mathcal{F}(\mathcal{T}\mathbf{X}) = \mathcal{T}\mathcal{F}(\mathbf{X})$, this property aids in training within only "canonical" space $\overline{\mathcal{T}}\mathcal{F}(\mathbf{X})$, while generalizing to any $\mathcal{T}\mathcal{F}(\mathbf{X})$. Gauge-equivariant neural networks extend 2D convolutions to manifolds by "shifting" kernels across tangent frames to achieve gauge symmetry equivariance [17]. Due to the high computational cost, the continuous $\mathrm{SO}(3)$ space is approximated using the icosahedron's 60 rotational symmetries. EPN [14] further extended this idea to point clouds and $\mathrm{SE}(3)$ by introducing separable convolutional layers that handle rotations and translations independently.

Formally, we discretize $\mathrm{SO}(3)$ into a finite rotation group $\mathcal{G}$ with $|\mathcal{G}| = 60$, where each element $\mathbf{g}_j$ corresponds to a rotation from the icosahedral group. As shown in Fig. 4, a continuous rotation amounts to a specific permutation of $\mathcal{G}$. Following prior work [22], we use this rotation group and its permutation operator to approximate the $\mathrm{SO}(3)$ rotation.

Given an input point cloud $\mathbf{X}$, the EPN [14] network produces an $\mathrm{SO}(3)$-equivariant feature $\mathcal{F} \in \mathbb{R}^{N\times O\times C}$, where N points and $O = 60$ group elements in $\mathcal{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_O\}$ each have a feature vector of size C. Mean pooling over the group dimension yields an invariant feature $\overline{\mathcal{F}} \in \mathbb{R}^{N\times C}$. These point-wise features capture localized, piecewise geometric information of the point cloud surface. In ArtEq [22], they are aggregated by body parts to estimate articulated pose, whereas ETCH predicts tightness vectors mapping outer cloth to the inner body instead, as illustrated in Fig. 4.

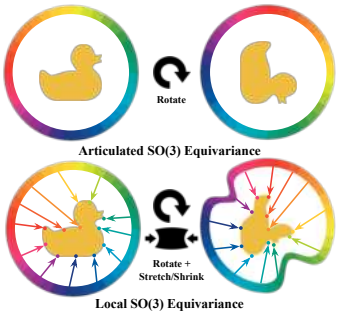

Figure 4. **SO(3) Equivariant Pose vs. Tightness.** Rainbow circle is the feature $\mathcal{F}(\mathbf{X})$, for articulated SO(3)-equiv, $\mathcal{T}$ denotes approximate rigid transformation of body part, while for our case, where the clothing roughly deforms with human poses, it refers to the tightness vector rotation.

3.2. Data Preparation

The key step is to establish the dense correspondence mapping $\phi : \mathbf{X} \rightarrow \mathbf{Y}$, between outer cloth points $\mathbf{X}$ and inner body points $\mathbf{Y}$. See mapping details in **"Anchor vs. Scatter Points"** at *Sup.Mat.*'s Sec. R.1. Given each established pointwise correspondence $\phi_i : \mathbf{x}_i \rightarrow \mathbf{y}_j$, $\mathbf{m}_k$ is the geodesic-closest marker of $\mathbf{y}_j$, all the other elements, including the tightness vector $\mathbf{v}_i$, label l_i, and confidence c_i, will be derived through:

$$\begin{aligned} l_i &= k, \quad \mathbf{v}_i = \mathbf{y}_j - \mathbf{x}_i \\ c_i &= \exp(-\lambda \times g(\mathbf{m}_k, \mathbf{y}_j; \mathcal{S}_\mathbf{Y})) \end{aligned} \tag{3}$$

where λ is the rate parameter of the exponential distribution to shrink it within the interval $[0, 1)$, and $g(\mathbf{m}_i, \mathbf{y}_j; \mathcal{S}_\mathbf{Y})$ is the geodesic distance defined on the inner surface $\mathcal{S}_\mathbf{Y}$.

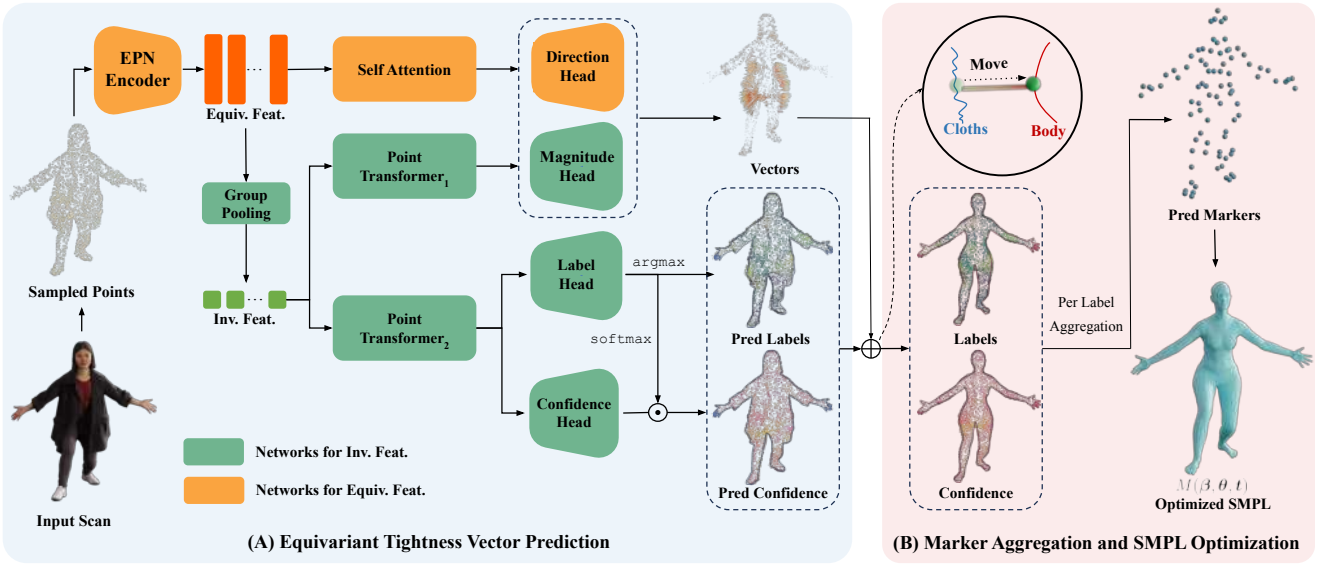

Figure 5. **ETCH Pipeline: 1) Equivariant Tightness Vector Prediction**, which takes the sampled points $\mathbf{X}$ as input, and estimates the tightness directions $\mathbf{D}$ via equivariant features $\mathbf{f}^{\text{equiv}}$ (Sec. 3.3), along with the tightness magnitudes $\mathbf{B}$, labels $\mathbf{L}$, and confidences $\mathbf{C}$ via invariant features $\mathbf{f}^{\text{inv}}$ (Sec. 3.4). With these ingredients, in **2) Marker Aggregation and SMPL Optimization**, the points move inward along the tightness vectors, forming body-shaped point clouds. These points are weighted and aggregated (Sec. 3.5), based on their labels and confidences, to produce final markers for SMPL fitting.

3.3. Tightness Vector Prediction

"Tightness vector $\mathbf{v}_i$" is the fundamental component of our system. Other components, including label $\mathbf{l}_i$ and confidence $\mathbf{c}_i$, are both derived from $\mathbf{v}_i$, as Eq. (3) shows. The tightness vector comprises two components: direction $\mathbf{d}_i$ and magnitude b_i, *i.e.* $\mathbf{v}_i = b_i\mathbf{d}_i$. The direction highly correlates with human articulated poses, thus it is learned with local approximate SE(3) equivariance, ensuring the tightness vector consistently maps the cloth surface to the body surface. While the magnitude mainly reflects clothing displacements, highly correlating with clothing types and body regions; thus, it is learned from the invariance features, as illustrated in Fig. 5.

Direction Prediction. To obtain the direction $\mathbf{d}_i$ while preserving the equivariance feature $\mathbf{f}_i^{\text{equiv}} = \mathcal{F}_{\text{EPN}}(\mathbf{X})_i \in \mathbb{R}^{O\times C}$, we treat the pointwise features $\mathbf{f}_i^{\text{equiv}}$ independently, where $\mathcal{F}_{\text{EPN}}$ denotes the EPN network [14], and $O = 60$ is the dimension of the rotation group $\mathcal{G}$. Specifically, we use a self-attention network $\mathcal{F}_{\text{self-attn}}$ to process the feature over the rotation group dimension (O), to ensure that each group element feature is associated with a group element $\mathbf{g}_j$ (which is a rotation in the discretized SO(3) group):

$$
\begin{aligned}
&\mathbf{f}_i^{\text{equiv}} = \mathcal{F}_{\text{EPN}}(\mathbf{X})_i \\
&w_{ij} = \mathcal{F}_{\text{mlp}}(\mathcal{F}_{\text{self-attn}}(\mathbf{f}_i^{\text{equiv}})_j), \quad j \in \{0, 1, \ldots, O-1\}, \\
&\hat{\mathcal{R}}_i = \mathcal{U}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(\mathcal{U}\mathcal{V}^T) \end{bmatrix}\mathcal{V}^T, \quad \mathcal{U}\mathcal{D}\mathcal{V}^T = \sum_{j=1}^{O} w_{ij}\mathcal{R}_j.
\end{aligned} \tag{4}
$$

where $\mathcal{F}_{\text{mlp}}$ is the direction head parameterized with an MLP network, $\mathcal{R}_j$ is the rotation matrix for $\mathbf{g}_j$, and $\mathcal{D}$ is a diagonal matrix in the matrix decomposition. Finally, $\hat{\mathcal{R}}_i$ is derived, which is the rotation matrix for $\mathbf{x}_i$. To transform from rotation to vector field, we multiply $\hat{\mathcal{R}}_i$ by a unit vector $\mathbf{v}_s$ (*e.g.* $\mathbf{v}_s = \begin{bmatrix}0 & 0 & 1\end{bmatrix}$) to obtain the final $\hat{\mathbf{d}}_i$, *i.e.*, $\hat{\mathbf{d}}_i = \hat{\mathcal{R}}_i\mathbf{v}_s^\top$.

Magnitude Prediction. To obtain $\hat{\mathbf{B}} = \{b_i \in \mathbb{R}\}_N$, we first aggregate (mean pooling) the feature over the rotation group dimension (O) for all points $\mathbf{X}$ to obtain the invariant feature $\mathbf{f}^{\text{inv}} = \overline{\mathcal{F}}_{\text{EPN}}(\mathbf{X}) \in \mathbb{R}^{N\times C}$. Instead of treating pointwise features $\mathbf{f}_i^{\text{inv}} = \overline{\mathcal{F}}_{\text{EPN}}(\mathbf{X})_i \in \mathbb{R}^C$ independently, we choose *Point Transformer* [70] $\mathcal{F}_{\text{PT-1}}$ to capture the contextual information outside this point.

$$
\hat{\mathbf{B}} = \mathcal{F}_{\text{Mag}}(\mathcal{F}_{\text{PT-1}}(\mathbf{f}^{\text{inv}}, \mathbf{X}; \delta)). \tag{5}
$$

where $\delta = \Theta(\mathbf{x}_i - \mathbf{x}_j)$ is the learned position embedding, parameterized with MLP Θ, which encodes the relative positions between point pairs $\{\mathbf{x}_i, \mathbf{x}_j\}$, and $\mathcal{F}_{\text{Mag}}$ is the magnitude head, which is a MLP.

3.4. Label and Confidence Prediction

Along the estimated tightness vectors $\hat{\mathbf{V}}$, sampled points $\mathbf{X}$ shoot the inner body, where we will aggregate the predicted inner points, with their estimated marker label $\hat{\mathbf{L}}$ and confidence $\hat{\mathbf{C}}$, to obtain the final sparse markers $\hat{\mathbf{M}}$. Notably, here we also use *Point Transformer*, denoted as $\mathcal{F}_{\text{PT-2}}$, since estimating labels and confidences is essentially a segmentation task, and the contextual information is crucial for distinguishing the symmetrical body parts (*i.e.*, hands, feet).

Label. Point Transformer $\mathcal{F}_{\text{PT-2}}$ and $\mathcal{F}_{\text{Label}}$ takes $\overline{\mathcal{F}}_{\text{EPN}}(\mathbf{X}) \in \mathbb{R}^{N\times C}$ with position $\mathbf{X} \in \mathbb{R}^{N\times 3}$ as input and outputs $\mathcal{P} \in \mathbb{R}^{N\times K}$, here $K = 86$, with $\mathcal{P}(\mathbf{x}_i, \mathbf{m}_k)$ representing the probability of point $\mathbf{x}_i$ belonging to marker $\mathbf{m}_k$:

$$\begin{aligned} \mathcal{P}(\mathbf{X}) &= \texttt{softmax}(\mathcal{F}_{\text{Label}}(\mathcal{F}_{\text{PT-2}}(\mathbf{f}^{\text{inv}}, \mathbf{X}; \delta))), \\ \hat{\mathbf{L}} &= \texttt{argmax}(\mathcal{F}_{\text{Label}}(\mathcal{F}_{\text{PT-2}}(\mathbf{f}^{\text{inv}}, \mathbf{X}; \delta))). \end{aligned} \tag{6}$$

Confidence. Following LoopReg [8], we use Group Convolution $\mathbf{C}_k$ for each marker $\mathbf{m}_k$, and soft aggregation to compute the confidence of each tightness vector:

$$\begin{aligned} \beta_{ik} &= \mathcal{F}_{\text{Conf}}(\mathcal{F}_{\text{PT-2}}(\mathbf{f}^{\text{inv}}, \mathbf{X}; \delta))_i * \mathbf{C}_k, \\ \hat{c}_i &= \sum_{k=1}^{K} \mathcal{P}(\mathbf{x}_i, \mathbf{m}_k)\beta_{ik}. \end{aligned} \tag{7}$$

where both $\mathcal{F}_{\text{Conf}}$ and $\mathcal{F}_{\text{Label}}$ are MLPs, $\mathcal{P}(\mathbf{X}) \in \mathbb{R}^{N\times K}$.

Now we have all the ingredients for cooking, the final loss $\mathcal{L}$ is formulated as follows:

$$\begin{aligned} \mathcal{L} &= w_d\mathcal{L}_d + w_b\mathcal{L}_b + w_l\mathcal{L}_l + w_c\mathcal{L}_c, \\ \mathcal{L}_d &= \sum_{i=1}^{N} \frac{\hat{\mathbf{v}}_i \cdot \mathbf{v}_i}{\|\hat{\mathbf{v}}_i\|\|\mathbf{v}_i\|}, \quad \mathcal{L}_b = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{b}}_i - \mathbf{b}_i)^2, \\ \mathcal{L}_l &= -\frac{1}{N}\sum_{i=1}^{N}\log(\mathcal{P}(\mathbf{x}_i, \mathbf{m}_{k=l_i})), \quad \mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{c}}_i - \mathbf{c}_i)^2. \end{aligned} \tag{8}$$

where w_d, w_b, w_l, w_c separately represents the weighted factor of losses of direction, magnitude, label and confidence, see more technical details in *Sup.Mat.*'s Sec. R.1.

3.5. Marker Aggregation and SMPL Regression

Marker Aggregation. Along the predicted tightness vector $\hat{\mathbf{v}}_i$, the outer points $\mathbf{x}_i$ are shot towards inner points $\hat{\mathbf{y}}_i$, via $\hat{\mathbf{y}}_i = \mathbf{x}_i + \hat{\mathbf{v}}_i$. Then the body markers are obtained through weighted aggregation as follows:

$$\hat{\mathbf{m}}_k = \frac{\sum_{i=1}^{m} \hat{\mathbf{y}}_i^{\hat{l}_i=k} \cdot (\hat{c}_i^{\hat{l}_i=k})^\alpha}{\sum_{i=1}^{m} (\hat{c}_i^{\hat{l}_i=k})^\alpha} \tag{9}$$

where only top-m points with the highest confidences among those where the $\hat{l}_i$ coincides with k are aggregated, α further amplifies their influence, enhancing results in practice. With these aggregated markers, SMPL parameters are optimized:

$$\min_{\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{t}} \sum_{k=1}^{K} \|\tilde{\mathbf{m}}_k - \hat{\mathbf{m}}_k\|^2 \tag{10}$$

, where $\tilde{m}_k$ denotes the markers on current SMPL estimate. Specifically, we use a damped Gauss-Newton optimizer based on the Levenberg–Marquardt algorithm [53].

4. Experiments

4.1. Datasets.

We select CAPE [39] and 4D-Dress [59], both featuring pre-captured underlying bodies, and a wide range of human poses, shapes, and clothing variations. Unlike 4D-Dress, CAPE has tighter-fitting clothing and less dynamics. Specifically, CAPE [39] contains 15 subjects with different body shapes; we split them as 4:1 to evaluate the robustness against "body shape and garment variations". Following NICP [41], we subsample by factors of 5 and 20 for training and validation sets, resulting in 26,004 training frames and 1,021 validation frames. For 4D-Dress [59], which contains 32 subjects with 64 outfits across over 520 motion sequences, we use the official split which selects 2 sequences per outfit to evaluate the robustness against "body pose and clothing dynamics variations". After subsampling by factors of 1 and 10 for training and validation respectively, we obtain 59,395 training frames and 1,943 validation frames. All the models are trained and tested with the above setting.

4.2. Metrics.

Distance Metrics. Three metrics are included: 1) vertex-to-vertex (V2V) distance in cm, between the correspondence vertices of our fitted and provided ground-truth SMPL bodies, 2) joint position error (MPJPE) in cm, measuring the Euclidean error of J SMPL joints, and 3) bidirectional Chamfer Distance in cm, between the predicted inner points (w/o SMPL fitting) and ground-truth SMPL bodies, a metric specifically designed to evaluate tightness-aware methods. Tabs. 1 and 2 and *Sup.Mat.*'s Tab. R.2 use these distance metrics, with lower values indicating more accurate body fits.

Directional Metrics. We ablate the effectiveness of Equivariance features by measuring the correctness of tightness vector direction $\mathbf{d}_i$, as this is the only element regressed from equivariance features. We calculate the angular error between the estimated and ground-truth tightness directions (detailed in Sec. 3.2). Specifically, in Tab. 3, we report both the mean and median cosine distance, with lower numbers indicating more accurate tightness directions.

4.3. Quantitative Results.

We compare our method with multiple SOTA baselines, spanning from tightness-aware approaches, *i.e.*, IPNet [7] and PTF [58], to tightness-agnostic ones, *i.e.*, NICP [41] and ArtEq [22], as shown in Tab. 1.

Baselines. ETCH achieves superior performance across all datasets and metrics. In particular, on CAPE, among all the competitors, our approach reduces the V2V error by $4.6\% \sim 36.5\%$ and MPJPE by $31.3\% \sim 51.9\%$; on 4D-Dress, the improvement is even more significant with a $16.7\% \sim 59.2\%$ decrease in V2V error and $32.6\% \sim 69.5\%$

Methods	CAPE [39]			4D-Dress [59]		
	V2V ↓	MPJPE ↓	CD ↓	V2V ↓	MPJPE ↓	CD ↓
Tightness-agnostic						
NICP [41]	1.726	1.343	-	4.754	3.654	-
ArtEq [22]	2.200	1.557	-	2.328	1.657	-
Tightness-aware						
IPNet [7]	2.593	1.917	1.110	3.826	2.625	1.262
PTF [58]	2.036	1.497	1.219	2.796	2.053	1.239
Ours	**1.647**	**0.922**	**1.019**	**1.939**	**1.116**	**1.065**

Table 1. **Quantitative Comparison with SOTAs.** ETCH clearly outperforms SOTAs, whether tightness-agnostic or -aware, in both CAPE and 4D-Dress across all metrics. In 4D-Dress-MPJPE, it surpasses the ArtEq by nearly 32.6%. Notably, for a fair comparison, no post-refinement is introduced to NICP [41] here, see NICP w/ post-refinement at *Sup.Mat.*'s Tab. R.2.

in MPJPE. Among tightness-aware methods (*i.e.*, IPNet, PTF and Ours), under bidirectional Chamfer Distance, our method achieves $8.2\% \sim 16.4\%$ improvement on CAPE and $14.0\% \sim 15.6\%$ on 4D-Dress between the predicted inner points/meshes (w/o SMPL fitting) and ground-truth SMPL bodies. We attribute this leading edge to our innovative design of "tightness vector", which effectively disentangles clothing and body layers, simplifying the fitting of loosely clothed humans into a bare-body fitting problem. This makes our method more robust to variations of bodies (shapes, poses) and clothing (types, fitness, dynamics).

It is worth noting that among optimization-based methods (*i.e.*, NICP, IPNet, and PTF), additional priors, regularizers, or post-refinement are incorporated during SMPL optimization, such as pose priors (*e.g.*, VPoser [45], GMM [11]), shape regularizer to encourage $\boldsymbol{\beta}$ to be small, inner-to-body point-mesh distance regularizer (*i.e.*, PTF, IPNet), and "Chamfer-based Post-Refinement" of NICP † in *Sup.Mat.*'s Tab. R.2. In contrast, our method optimizes SMPL using only the 86 markers (Sec. 3.5) directly predicted by ETCH as optimization targets, ***without bells and whistles*** (*i.e.*, pose priors, and geometric constraints), yet still achieves substantial performance gains, further validating our key design. In addition, the body fitting error could arise from either pose or shape errors; joint error MPJPE in Tab. 1 only confirms pose accuracy, so we evaluate "shape accuracy" in *Sup.Mat.*'s Fig. R.3, showing average 49.9% improvement compared to other optimization-based approaches.

4.4. Qualitative Results.

In *Sup.Mat.*'s Fig. R.4, each row illustrates a challenging case: A, B and C focus on difficult poses, while D, E, F and G highlight loose garments with large dynamic deformations. Our method achieves pixel-level body alignment, showing its superior robustness to various poses, shapes, and garments. More analysis is detailed in the caption of *Sup.Mat.*'s Fig. R.4. *Sup.Mat.*'s Fig. R.1 shows ETCH trained on 4D-Dress generalize well on unseen THuman2.1 scans and HuGe100K generated 3D humans [74]. Please check out more video results in `website`.

	Ablation Settings							CAPE [39]		4D-Dress [59]	
	Tightness		Correspondence		Features for Direction $\mathbf{d}_\perp$						
Settings	direction	scalar	dense	markers	Inv	XYZ	Equiv	V2V ↓	MPJPE ↓	V2V ↓	MPJPE ↓
Ours	✓	✓	✗	✓	✗	✗	✓	**1.647**	**0.922**	**1.939**	**1.116**
A.	✓	✓	✗	✓	✗	✓	✗	1.661	0.925	2.033	1.134
B.	✓	✓	✗	✓	✓	✓	✗	1.663	0.926	2.307	1.314
C.	✓	✓	✓	✗	✗	✗	✓	1.909	1.451	2.285	1.466
D.	✗	✓	✓	✗	✗	✗	✓	1.777	1.342	2.410	1.608
E.	✓	✓	✗	✓	✓	✗	✗	1.888	1.101	2.842	2.014

Table 2. **Ablation Study of ETCH.** Please check Sec. 4.5 for more in-depth analysis, and Tab. 3 and Fig. 7 to explore OOD generalization of equivariance features. For simplicity, "Inv" denotes Invariance Features, "Equiv" denotes Equivariance Features, "XYZ" denotes XYZ-Positions. The full-featured ETCH is referred to as "Ours", while variants are labeled "Ours-X". Ours-A and Ours-B replace equivariance features with xyz-positions and/or invariance features, while Ours-E only uses invariance features. Although invariant feature is inherently incompatible with Direction prediction, we add Ours-E here for more comprehensive comparison. Ours-C and Ours-D use dense correspondence, with Ours-D removing the direction term to assess its necessity.

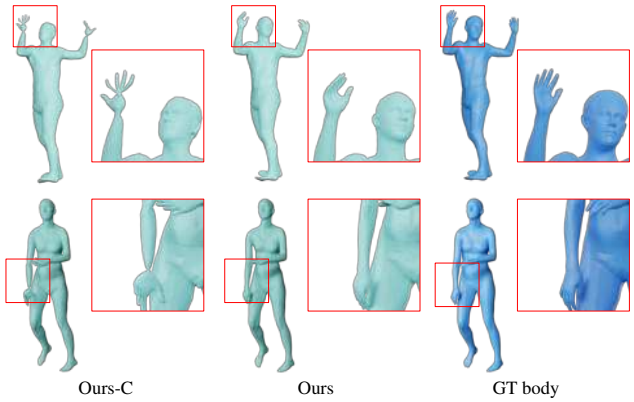

Figure 6. **Sparse Markers vs. Dense Correspondence.** Incorrect rotation angles and flipped configurations are evident in the head and hand regions of the first row, and in the forearm and hand regions of the second row. More quantitative results in Tab. 2 and its analysis in Sec. 4.5.

4.5. Ablation Studies

We conduct ablation studies to explore strategies like "Sparse Marker vs. Dense Correspondence" (Tab. 2 and Fig. 6) and validate design choices, such as the Tightness Scalar in "Tightness Vector vs. Scalar" (Tab. 2), Equivariance features in "w/ Equiv vs. w/o Equiv" under full training set and one-shot settings (Tab. 3 and Fig. 7), the influence of chamfer-based post-reinforcement (*Sup.Mat.*'s Tab. R.2), and shape accuracy (*Sup.Mat.*'s Fig. R.3). The full-featured ETCH is referred to as "Ours", while variants are labeled "Ours-X". Ours-A and Ours-B replace equivariance features with xyz-positions and/or invariance features, Ours-E only uses invariance features. Although invariance feature is inherently incompatible with Direction prediction, we add Ours-E here for more comprehensive comparison. Ours-C and Ours-D use dense correspondence, with Ours-D removing the direction term to assess its necessity.

Sparse Marker vs. Dense Correspondence. In Tab. 2, the comparison between Ours and Ours-C demonstrates the strength of our sparse marker design. To clarify, we first use tightness vectors to find the mapping between the scan

Settings	Features for Direction $\mathbf{d}_i$			CAPE [39]		4D-Dress [59]	
	Inv	XYZ	Equiv	Mean ↓	Median ↓	Mean ↓	Median ↓
Ours	✗	✗	✓	**0.616**	**0.0535**	**0.919**	**0.315**
A.	✗	✓	✗	0.763	0.527	0.986	0.988
B.	✓	✓	✗	0.664	0.299	0.963	0.961

Table 3. **Equivariance Generalizes well in One-shot Settings** For simplicity but aligned with Tab. 1, "Inv" denotes Invariance Features, "Equiv" denotes Equivariance Features, "XYZ" denotes XYZ-Positions. Fig. 7 shows the directional error (left), and predicted inner body points (right).

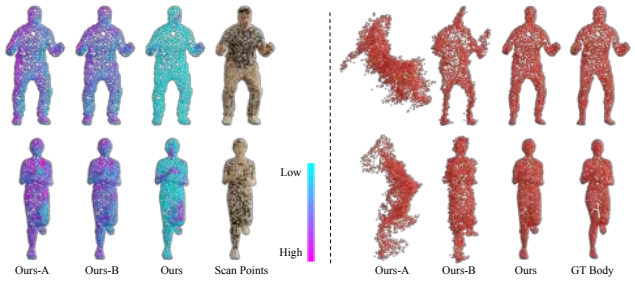

Figure 7. **Equivariance for OOD.** We test multiple variants of ETCH on one-shot settings to illustrate the out-of-distribution (OOD) generalization of equivariant features, by replacing the equivariant features (Ours) with xyz-positions (Ours-A) or xyz-positions+invariance (Ours-B). The visualization shows scan points colored by angular error (left) and predicted inner points (right). The error bar indicates cosine error of tightness vector.

mesh and the SMPL surface. Then, each position on the posed SMPL surface corresponds to a position in canonical SMPL space, which serves as the **Dense Correspondence**. For Ours, by mapping dense points to sparse markers and aggregating them with confidence, we create a voting strategy that enhances robustness against low-confidence outliers. This approach is more effective than dense prediction, which can struggle with outliers and local minima. Our method achieves significant improvements: 13.7% lower V2V error and 36.5% reduced MPJPE on CAPE, along with an error decrease of 15.1% and 23.9% on 4D-Dress. These consistent gains validate our use of sparse markers over dense correspondences. Figure 6 shows their differences; incorrect dense correspondences can misguide optimization, leading to skewed body part rotations (*e.g.*, hands, forearms, head), while our sparse marker strategy remains robust through its weighted voting mechanism.

Tightness Vector vs. Tightness Scalar. In Tab. 2, we compare the tightness vector (Ours-C) and scalar (Ours-D) under a dense correspondence setting. The model performs better without the tightness vector on the CAPE dataset, while the opposite is true for 4D-Dress. For tight clothing with minimal cloth-to-body tightness, the scalar-only (Ours-D) works well for SMPL optimization. However, incorrect direction estimates – like being flipped – can disrupt fitting due to sensitivity to outliers. For loose garments, the tightness scalar lacks cues for inner points, making accurate direction crucial. Thus, when fit styles are uncertain, especially with loose clothing, introducing the tightness vector is advisable.

w/ Equiv vs. w/o Equiv. In Tabs. 2 and 3, we assess how equivariant features affect direction prediction at two scales:

1) Full Training Set: Compare our method with Ours-A (xyz positions) and Ours-B (xyz+invariance) as inputs to the *Point Transformer*. All three methods perform similarly on CAPE, but our approach has a slight edge. The gap widens on the 4D-Dress dataset, showing improvements of 4.6% $\sim$ 16.0% in V2V and 1.6% $\sim$ 15.1% in MPJPE. This suggests that equivariance is the key to addressing hard cases, such as loose garments and significant pose variations. According to hypothesis proposed in Sec. 3.3, invariance feature inherently conflicts with Direction prediction. The poor performance of Ours-E further demonstrates this point.

2) One-shot Settings: To further illustrate the out-of-distribution (OOD) generalization of equivariant features, we randomly sampled one frame from each sequence to create minimal training sets (0.66% of the 4D-Dress train split and 2.1% of the CAPE train split). We train Ours, Ours-A, and Ours-B on them and evaluate on the same validation sets as in Tab. 2. In Fig. 7, the visualization shows scan points colored by angular error (left) and predicted inner points (right). Our method shoots inner points in the correct directions in most areas and the inner points closely approximate the inner body shape, while Ours-A and Ours-B fail completely. Tab. 3 presents mean and median angular errors (Sec. 4.2) between predicted and ground truth directions, highlighting our method's advantages, with notable median angular error improvements of 82.1% $\sim$ 89.8% on CAPE and 67.2% $\sim$ 68.1% on 4D-Dress. These results demonstrate that equivariant features enable robust direction prediction and strong OOD generalization with limited data.

5. Conclusion

ETCH generalizes the fitting of the *underlying body* to diverse *clothed humans* through a novel framework powered by Equivariant Tightness Vector and Marker-Aggregation Optimization. We are neither the first to disentangle the clothing layer from clothed scans [7, 58] nor the pioneers in modeling tightness [15, 25], but we are the first to do so using equivariant vectors, significantly outperforming previous works across various datasets and metrics. While there is still room for improvement as detailed in *Sup.Mat.*'s Sec. R.3, ETCH paves the way for truly generalizable cloth-to-body fitting that is robust to any body pose and shape, garment type, and non-rigid dynamics, achieving fundamental improvements. We believe ETCH will reshape new perspectives on not only body fitting, but also garment refitting, real-to-sim, and biomechanics analysis. However, this technique could be misused by the porn industry and pose a threat to human privacy; therefore, we will release the packages solely for non-commercial scientific research purposes under a strict license to regulate user behavior.

Disclosure While MJB is a co-founder and Chief Scientist at Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society.

Acknowledgments. We thank *Marilyn Keller* for the help in Blender rendering, *Brent Yi* for fruitful discussions, *Ailing Zeng* and *Yiyu Zhuang* for HuGe100K dataset, *Jingyi Wu* and *Xiaoben Li* for their help during rebuttal, and the members of *Endless AI Lab* for their help and discussions. This work is funded by the Research Center for Industries of the Future (RCIF) at Westlake University, the Westlake Education Foundation. *Yuliang Xiu* also received funding from the Max Planck Institute for Intelligent Systems.

References

[1] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. *Transactions on Graphics (TOG)*, 22(3): 587–594, 2003. 3

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 408–416, 2005. 3

[3] Dimitrije Antić, Garvita Tiwari, Batuhan Ozcomlekci, Riccardo Marin, and Gerard Pons-Moll. CloSe: a 3D clothing segmentation dataset and model. In *International Conference on 3D Vision (3DV)*, pages 591–601. IEEE, 2024. 3

[4] Matan Atzmon, Jiahui Huang, Francis Williams, and Or Litany. Approximately Piecewise E (3) Equivariant Point Networks. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[5] A. Balan and M. J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision (ECCV)*, pages 15–29, Marseilles, France, 2008. Springer-Verlag. 2

[6] Bharat Bhatnagar, Ilya Petrov, and Xianghui Xie. RVH Mesh Registration. https://github.com/bharat-b7/RVH_Mesh_Registration, 2022. 3

[7] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 2, 3, 6, 7, 8, 1

[8] 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. *Conference on Neural Information Processing Systems (NeurIPS)*, 33:12909–12922, 2020. 3, 6

[9] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 1

[10] Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3794–3801, 2014. 3

[11] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. 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. 7

[12] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic FAUST: Registering human bodies in motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6233–6242, 2017. 3

[13] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 3

[14] Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. Equivariant point network for 3d point cloud analysis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14514–14523, 2021. 3, 4, 5

[15] Xin Chen, Anqi Pang, Wei Yang, Peihao Wang, Lan Xu, and Jingyi Yu. TightCap: 3D human shape capture with clothing tightness field. *Transactions on Graphics (TOG)*, 41(1):1–17, 2021. 2, 3, 8

[16] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992. 3

[17] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *International conference on Machine learning*, pages 1321–1330. PMLR, 2019. 4

[18] Lu Dai, Liqian Ma, Shenhan Qian, Hao Liu, Ziwei Liu, and Hui Xiong. Cloth2Body: Generating 3D Human Body Mesh from 2D Clothing. In *International Conference on Computer Vision (ICCV)*, pages 15007–15017, 2023. 1

[19] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *International Conference on Computer Vision (ICCV)*, pages 12200–12209, 2021. 3

[20] Congyue Deng, Jiahui Lei, Bokui Shen, Kostas Daniilidis, and Leonidas J. Guibas. Banana: Banach Fixed-Point Network for Pointcloud Segmentation with Inter-Part Equivariance. *ArXiv*, abs/2305.16314, 2023. 3

[21] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(6):7157–7173, 2022. 3

[22] 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)*, pages 7977–7988, 2023. 2, 3, 4, 6, 7, 1

[23] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3d roto-translation equivariant attention networks. *Conference on Neural Information Processing Systems (NeurIPS)*, 33:1970–1981, 2020. 3

[24] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal Human Parsing via Graph Transfer Learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[25] Peng Guan, Oren Freifeld, and Michael J Black. A 2D human body model dressed in eigen clothing. In *European Conference on Computer Vision (ECCV)*, pages 285–298. Springer, 2010. 2, 8, 1

[26] Nils Hasler, Carsten Stoll, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Estimating body shape of dressed humans. *Computers & Graphics (CG)*, 33(3): 211–216, 2009. IEEE International Conference on Shape Modelling and Applications 2009. 2

[27] David A Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conference on Computer Vision (ECCV)*, pages 242–255. Springer, 2012. 3

[28] Boyan Jiang, Yinda Zhang, Xingkui Wei, Xiangyang Xue, and Yanwei Fu. H4d: Human 4d modeling by learning neural compositional representation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 19355–19365, 2022. 3

[29] Haiyong Jiang, Jianfei Cai, and Jianmin Zheng. Skeleton-aware 3D human shape reconstruction from point clouds. In *International Conference on Computer Vision (ICCV)*, pages 5431–5441, 2019. 3

[30] Maria Korosteleva and Olga Sorkine-Hornung. GarmentCode: Programming Parametric Sewing Patterns. *Transactions on Graphics (TOG)*, 42(6), 2023. SIGGRAPH ASIA 2023 issue. 1

[31] Maria Korosteleva, Timur Levent Kesdogan, Fabian Kemper, Stephan Wenninger, Jasmin Koller, Yuhan Zhang, Mario Botsch, and Olga Sorkine-Hornung. GarmentCodeData: A Dataset of 3D Made-to-Measure Garments With Sewing Patterns. In *European Conference on Computer Vision (ECCV)*, 2024.

[32] Boqian Li, Xuan Li, Ying Jiang, Tianyi Xie, Feng Gao, Huamin Wang, Yin Yang, and Chenfanfu Jiang. GarmentDreamer: 3DGS Guided Garment Synthesis with Diverse Geometry and Texture Details. In *International Conference on 3D Vision (3DV)*, 2025. 1

[33] Hao Li, Robert W. Sumner, and Mark Pauly. Global Correspondence Optimization for Non-Rigid Registration of Depth Scans. *Computer Graphics Forum (CGF)*, 27(5), 2008. 3

[34] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. PSHuman: Photorealistic Single-view Human Reconstruction using Cross-Scale Diffusion. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 1

[35] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! Text to Animatable Digital Avatars. In *International Conference on 3D Vision (3DV)*, 2024. 1

[36] Guanze Liu, Yu Rong, and Lu Sheng. Votehmr: Occlusion-aware voting network for robust 3d human mesh recovery from partial point clouds. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 955–964, 2021. 3

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *Transactions on Graphics (TOG)*, 2015. 1, 3, 4

[38] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and Shape Capture from Sparse Markers. *Transactions on Graphics (TOG)*, 33(6):220:1–220:13, 2014. 2

[39] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 4, 6, 7, 8, 0

[40] 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)*, pages 5442–5451, 2019. 3

[41] Riccardo Marin, Enric Corona, and Gerard Pons-Moll. NICP: Neural ICP for 3D Human Registration at Scale. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 6, 7, 0

[42] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On Self-Contact and Human Pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[43] Alexandros Neophytou and Adrian Hilton. A Layered Model of Human Body and Garment Deformation. In *International Conference on 3D Vision (3DV)*, pages 171–178, 2014. 3

[44] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in Geography Optimized for Regression Analysis. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[45] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[46] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67: 276–286, 2017. 3

[47] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: A model of dynamic human shape in motion. *Transactions on Graphics (TOG)*, 34(4):1–14, 2015. 3

[48] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D Clothing Capture and Retargeting. In *Transactions on Graphics (TOG)*, 2017. Two first authors contributed equally. 3

[49] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *International Conference on Computer Vision (ICCV)*, pages 4332–4341, 2019. 3

[50] Omri Puny, Matan Atzmon, Heli Ben-Hamu, Edward James Smith, Ishan Misra, Aditya Grover, and Yaron Lipman. Frame Averaging for Invariant and Equivariant Network Design. *ArXiv*, abs/2110.03336, 2021. 3

[51] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 1, 3

[52] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on

point sets in a metric space. *Conference on Neural Information Processing Systems (NeurIPS)*, 30, 2017. 1, 3

[53] Sam Roweis. Levenberg-marquardt optimization. *Notes, University Of Toronto*, 52:1027, 1996. 6

[54] Dan Song, Ruofeng Tong, Jian Chang, Xiaosong Yang, Min Tang, and Jian Jun Zhang. 3D Body Shapes Estimation from Dressed-Human Silhouettes. *Computer Graphics Forum (CGF)*, 35:147–156, 2016. 2

[55] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *International Conference on Computer Vision (ICCV)*, pages 6411–6420, 2019. 3

[56] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 3

[57] Kangkan Wang, Jin Xie, Guofeng Zhang, Lei Liu, and Jian Yang. Sequential 3D human pose and shape estimation from point clouds. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7275–7284, 2020. 3

[58] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally Aware Piecewise Transformation Fields for 3D Human Mesh Registration. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 6, 7, 8, 0

[59] Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4D-DRESS: A 4D Dataset of Real-world Human Clothing with Semantic Annotations. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 4, 6, 7, 8, 0

[60] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Conference on Neural Information Processing Systems (NeurIPS)*, 35:33330–33342, 2022. 3

[61] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4840–4851, 2024. 3

[62] Stefanie Wuhrer, Leonid Pishchulin, Alan Brunton, Chang Shu, and Jochen Lang. Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding (CVIU)*, 127:31–42, 2014. 2

[63] Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitrios Tzionas, and Michael J Black. PuzzleAvatar: Assembling 3D Avatars from Personal Albums. *Transactions on Graphics (TOG)*, 2024. 1

[64] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[65] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Estimation of Human Body Shape in Motion with Wide Clothing. In *European Conference on Computer Vision (ECCV)*, pages 439–454, Cham, 2016. Springer International Publishing. 2

[66] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. SynBody: Synthetic Dataset with Layered Human Models for 3D Human Perception and Modeling. In *International Conference on Computer Vision (ICCV)*, pages 20282–20292, 2023. 1

[67] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 4

[68] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Conference on Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3

[69] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, Accurate, Human Shape Estimation From Clothed 3D Scan Sequences. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[70] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021. 3, 5

[71] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D Human Reconstruction From a Single Image. In *International Conference on Computer Vision (ICCV)*, 2019. 3, 4

[72] Jianqi Zhong, Ta-Ying Cheng, Yuhang He, Kai Lu, Kaichen Zhou, A. Markham, and Niki Trigoni. Multi-body SE(3) Equivariance for Unsupervised Rigid Segmentation and Motion Estimation. *ArXiv*, abs/2306.05584, 2023. 3

[73] Boyao Zhou, Jean-Sébastien Franco, Federica Bogo, Bugra Tekin, and Edmond Boyer. Reconstructing human body mesh from point clouds by adversarial gp network. In *Asian Conference on Computer Vision (ACCV)*, 2020. 3

[74] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. IDOL: Instant Photorealistic 3D Human Creation from a Single Image. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 7, 0, 1

[75] Xingxing Zou, Xintong Han, and Waikeung Wong. Cloth4d: A dataset for clothed human reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12847–12857, 2023. 1

[76] Silvia Zuffi and Michael J Black. The stitched puppet: A graphical model of 3d human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3537–3546, 2015. 3