

ETCH-X: Robustify Expressive Body Fitting to Clothed Humans with Composable Datasets

Xiaoben Li^{1,2,3}, Jingyi Wu^{2,4}, Zeyu Cai^{2,5}, Siyuan Yu²,
Boqian Li², and Yuliang Xiu^{2*}

¹ Zhejiang University, Hangzhou, China

² Westlake University, Hangzhou, China

³ Shanghai Innovation Institute, Shanghai, China

⁴ Fudan University, Shanghai, China

⁵ Nanjing University, Nanjing, China

{lixiaoben, yusiyuan, xiuyuliang}@westlake.edu.cn,

jingyiwu23@fudan.edu.cn, caizeyu010612@gmail.com,

boqianlihuister@gmail.com [xiaobenli00.github.io/ETCH-X](https://github.com/xiaobenli00/ETCH-X)

Abstract. Human body fitting, which aligns parametric body models, such as SMPL, to raw 3D point clouds of clothed humans, serves as a crucial first step for downstream tasks like animation and texturing. An effective fitting method should be both **locally expressive** – capturing fine details such as hands and facial features – and **globally robust** to handle real-world challenges, including clothing dynamics, pose variations, and noisy or partial inputs. Existing approaches typically excel in only one aspect, lacking an all-in-one solution. We upgrade ETCH to ETCH-X, which leverages a tightness-aware fitting paradigm to filter out clothing dynamics (“undress”), extends expressiveness with SMPL-X, and replaces explicit sparse markers (which are highly sensitive to partial data) with implicit dense correspondences (“dense fit”) for more robust and fine-grained body fitting. Our disentangled “undress” and “dense fit” modular stages enable separate and scalable training on composable data sources, including diverse simulated garments (CLOTH3D), large-scale full-body motions (AMASS), and fine-grained hand gestures (InterHand2.6M), improving outfit generalization and pose robustness of both bodies and hands. Our approach achieves robust and expressive fitting across diverse clothing, poses, and levels of input completeness, delivering a substantial performance improvement over ETCH on both 1) seen data, such as 4D-Dress (MPJPE-All, **33.0% ↓**) and CAPE (V2V-Hands, **35.8% ↓**), and 2) unseen data, such as BEDLAM2.0 (MPJPE-All, **80.8% ↓**; V2V-All, **80.5% ↓**). Code and models will be released at [xiaobenli00.github.io/ETCH-X](https://github.com/xiaobenli00/ETCH-X).

Keywords: Clothed humans · Partial scans · 3D Body fitting · Hand pose estimation · Dense correspondences

* Corresponding author.

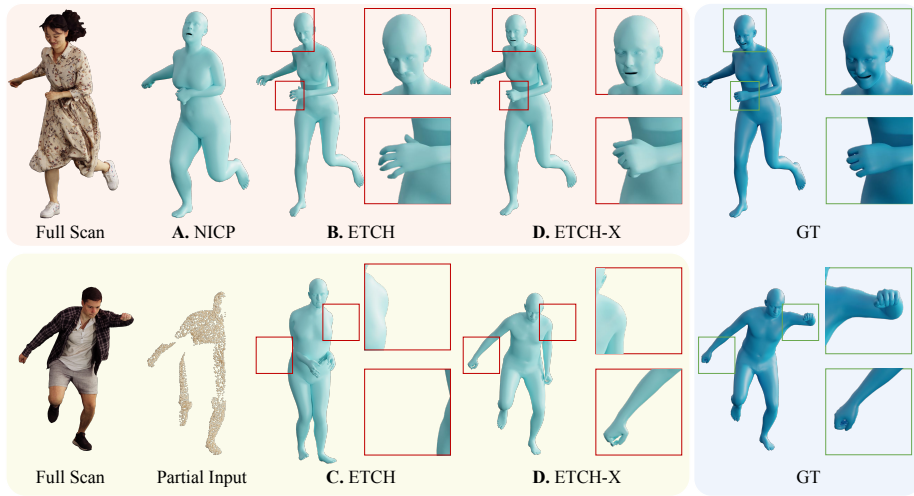


Fig. 1: **Strengths of ETCH-X.** While NIPC [31], which uses implicit dense correspondence but lacks tightness-aware undressing, consistently produces overweight bodies from clothed scans (A), ETCH [24], with tightness-aware undressing but sparse markers, fails to capture detailed body parts such as hands and face (B), and struggles with partial inputs due to missing markers (C). In contrast, our ETCH-X combines the strengths of both approaches, achieving robust and expressive fitting across diverse clothing, poses, and levels of input completeness (D).

1 Introduction

Humans are commonly captured as point clouds using 3D scanners or depth sensors. Such point clouds are often noisy, incomplete, and lack topological structure, preventing direct use in downstream tasks, such as shape analysis [1, 8], animation [28, 41], garment refitting [20, 25], and human interactions [22]. A crucial step to enable these applications is to align a parametric human body model (*e.g.*, SMPL-X [35], GHUM [54]) to the raw point cloud, producing a topologically consistent mesh with known correspondences. This process is commonly referred as “human body fitting”.

In particular, we focus on fitting the expressive parametric body model, *i.e.* SMPL-X [35], which includes detailed hand gestures and facial expressions, to raw point clouds of clothed humans. This is a challenging problem due to the large variation in clothing styles, body shapes, and poses, as well as the presence of noise and partial observations in the input point clouds. Even worse, the 3D clothed scans with perfectly aligned SMPL-X ground-truth are extremely scarce, and collecting such data is labor-intensive and costly.

Body fitting pipelines typically involve two steps: 1) establishing correspondences between the input clothed human point cloud and the body model template, like SMPL [27], via ICP (iterative closest point) or its variants [1, 13, 19, 36, 61], and 2) optimizing [5–7, 24, 31, 42, 50] or regressing [19] the model parameters to minimize the distance between corresponding points. The correspondence could

be dense [6, 31, 50] or sparse, such as inner keypoints [7, 42], surface markers [24], and part-based labels [5].

Dense correspondence is inherently ill-defined for clothed humans, as the outer clothing surface can deviate substantially from the underlying body, especially for loose or dynamic garments. This deviation introduces ambiguity and instability, since a single point on a loose T-shirt, for example, may correspond to multiple possible locations on the inner body, and these correspondences can change as the clothing deforms. While some recent works attempt to learn dense correspondences to align the parametric body model to the clothed surface [6, 31, 50], the resulting fitted bodies often appear unnatural, overweight, or biomechanically implausible, as illustrated in Fig. 1-A.

For sparse correspondence, several practical solutions exist. For instance, 2D keypoint estimators [10, 18] trained on in-the-wild images of clothed humans can provide reasonably accurate and robust 2D keypoints, even when the underlying body is occluded. Additionally, surface-based sparse markers can be weighted and aggregated (voting with confidence) from dense correspondence [24] to improve stability. However, inner keypoints capture only the skeleton, providing limited information about body shape (fat or slim). Surface markers are often too sparse to capture fine details such as hand gestures and facial expressions, as illustrated in Fig. 1-B, which are crucial for many applications, like human-object interaction. Moreover, with partial input point clouds, these sparse correspondences may be entirely absent, causing the fitting process to fail, see Fig. 1-C.

The limitations of sparse correspondence – namely, reduced expressiveness for fine-grained parts and vulnerability to partial inputs – can be mitigated by adopting dense correspondence. However, as discussed above, dense correspondence becomes inherently ambiguous in the presence of clothing. This raises a key question: how can we balance **local expressiveness** and **global robustness** in human body fitting? – *Undress first, then dense fit!*

However, this requires an accurate “undressing” operation, which is itself a challenging problem. Inspired by ETCH [24], which learns *SE(3) equivariant tightness vectors* to effectively disentangle clothing from the underlying body, we extend it as ETCH-X, by retaining the original tightness vector regressor while replacing its explicit sparse markers with implicit full-body dense correspondence [15, 31]. The tightness vector regressor is critical for robust undressing, as the learned tightness vectors are locally SE(3) equivariant, providing reliable cues even when portions of the input point cloud are missing. The implicit full-body correspondence, defined on the expressive SMPL-X [35] model, enables dense matching between undressed human scans and the body, capturing fine details such as hand gestures and facial features. Furthermore, the implicit representation is inherently robust to partial inputs – after training with partial augmentation, *fullset* SMPL-X anchor points can be queried at any location in the entire feature space. ETCH-X, therefore, achieves robustness to partial inputs, and expressiveness for fine details, see Fig. 1-D.

More importantly, such “undress first, then dense fit” paradigm echos the emerging trend of scaling efforts in computer vision [16, 43, 48]. Since the “undress”

and “dense fit” modules are disentangled, we can independently leverage 1) unlimited simulated garments (*i.e.*, CLOTH3D [3]) and 2) large-scale pose libraries (*i.e.*, AMASS [30] for body poses and InterHand2.6M [32] for hand poses) to train each module separately. This modular approach enables us to combine them for superior robustness in fitting in-the-wild clothed human scans. In other words, simulated garments enrich the diversity of clothing styles for tightness-aware undressing, while large-scale pose libraries enhance the generalization of implicit dense fitting to various bodies and hand gestures.

We conduct a comprehensive evaluation of ETCH-X against SOTA methods on both in-distribution datasets, such as CAPE [29] and 4D-Dress [51], as well as out-of-distribution data from BEDLAM2.0 [44]. ETCH-X consistently demonstrates superior performance in terms of expressiveness (*e.g.*, accurately capturing hand gestures and facial details) and robustness (*e.g.*, effectively handling partial inputs). To validate the effectiveness of the “*undress first, then dense fit*” paradigm, we benchmark against “*dense fit only*” approaches like NICP [31] and “*undress first, then sparse fit*” methods such as ETCH [24]. Additionally, we ablate two technical innovations: hand refinement by re-sampling, which produce better hand poses, and skin-aware tightness masking, which rectifies tightness vectors on skin regions to improve undressing performance. Finally, we conduct a scaling analysis on simulated garments, CLOTH3D, and body pose libraries, AMASS, highlighting the potential for future scaling towards truly generalizable human body fitting.

In summary, we upgrade ETCH [24] in three key aspects, with all % values indicating reduced error over ETCH:

- **More Expressive.** Replacing SMPL with SMPL-X, employing implicit dense correspondence, and introducing re-sampling based hand refinement, ETCH-X captures finer hands (V2V-35.8% ↓ on CAPE) and head (V2V-8.1% ↓ on 4D-Dress), which are crucial for contact-rich interactions.
- **More Robust.** By decoupling the “*undress*” and “*dense fit*” modules, ETCH-X is robust to diverse clothing styles and pose variations. The locality of the tightness vector, replacement of sparse markers with implicit dense correspondence, and partial augmentation further improve its robustness on partial inputs (V2V-72.5% ↓ on 4D-Dress) and unseen human captures (MPJPE-80.8% ↓ on BEDLAM2.0).
- **More Scalable.** ETCH-X seamlessly scales with large-scale 3D garments and pose libraries, both of which are easier to simulate or collect than real scans. This scalability further reduces fitting error (MPJPE-27.2% ↓ on BEDLAM2.0). These results underscore the effectiveness of our modular design and highlight the potential for future scaling towards truly generalizable human body fitting.

2 Related Work

Fitting human body models to point clouds is fundamental to many human-centric tasks. Over the years, a wide range of methods have been proposed to tackle this challenge. We analyze them from three key perspectives: optimization vs. learning,

tightness-agnostic vs. tightness-aware, and sparse vs. dense correspondence. We also clarify how ETCH-X is positioned within this landscape.

Optimization vs. Learning. Early optimization-based human body fitting methods typically rely on the ICP algorithm [13] or its variants [1, 36, 61]. Modern optimization-based approaches [4, 29, 34, 42, 55, 57, 59] often involve complex pipelines with multiple intermediate steps, such as pose estimation [10, 18], body segmentation [2, 21], and triangulation, each potentially introducing errors that can accumulate and degrade final accuracy. While optimization-based methods can achieve highly accurate results given precise correspondences, they are generally time-consuming, motivating the development of more efficient alternatives.

Learning-based methods leverage large-scale 3D human datasets [29, 30, 51] and deep neural networks [38, 39, 45, 52, 53, 56, 58] designed for point cloud processing. These approaches either provide good initialization for subsequent fitting [5, 6, 50], directly regress body meshes [37, 49, 60], or predict statistical body model parameters [19, 23, 26] in a feed-forward manner, offering much faster inference but sometimes less accuracy than optimization-based methods. To balance speed and accuracy, hybrid approaches first estimate sparse markers [24] or dense correspondences [31], and then refine body parameters via optimization. ETCH-X adopts this hybrid paradigm.

Tightness-agnostic vs. Tightness-aware. Many methods [6, 19, 31], optimization- or learning-based, fit the human body model directly to the input point cloud. This works well for tight clothing, but fails for loose clothing, where the true body shape can deviate significantly from the observed surface, leading to inaccurate fits. To address this, “tightness-aware” methods [5, 12, 24, 50] explicitly model clothing to recover more accurate body shapes. TightCap [12] uses a clothing tightness field—displacements from garment to body in UV space. IPNet [5] and PTF [50] jointly predict inner and outer body surfaces via double-layer occupancy. ETCH [24] encodes tightness as displacement vectors from the cloth surface to the underlying body. ETCH-X extends ETCH’s tightness vector by introducing skin-aware masking, setting tightness to zero on uncovered skin regions.

Sparse Correspondence vs. Dense Correspondence. Establishing correspondence is a crucial step in human body fitting, typically categorized as either “sparse” or “dense.” Sparse correspondence [19, 24] often relies on part-based feature aggregation, which provides some robustness to noise but struggle with incomplete input, as missing regions may result in lost markers. In contrast, explicit dense correspondence [5, 6, 31, 50] queries pointwise correspondence features from a learned implicit field. Although dense sampling can be computationally intensive, it is generally more robust to partial inputs—especially when combined with partial data augmentation—and better captures fine details. ETCH-X adopts the implicit dense correspondence strategy, following NICP [31], and extends it with re-sampling based hand refinement to enhance hand pose accuracy.

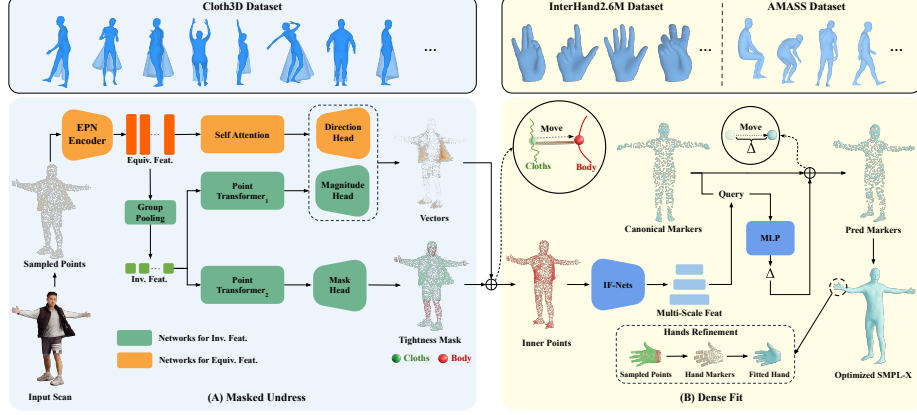


Fig. 2: **Two stages of ETCH-X: (A) Masked Undress, (B) Dense Fit.** In the Masked Undress stage, we take a clothed scan as input and compute the undressed body ($\hat{\mathbf{y}}_i = \mathbf{x}_i + \hat{\mathbf{l}}_i \hat{\mathbf{v}}_i$). In the Dense Fit stage, we implicitly learn the deforming field, which deforms the canonical SMPL-X into a posed one. Thanks to the decoupled design, the robustness to dynamic clothing and pose variations could be improved with simulated garments, *i.e.*, CLOTH3D [3], and pose libraries, *i.e.*, AMASS [30] for body poses and InterHand2.6M [32] for hand poses, respectively.

3 Method

As illustrated in Fig. 2, following the “undress first, then dense fit” paradigm, our method ETCH-X consists of two stages: 1) *masked undress*, which learns $SE(3)$ equivariant tightness vectors and skin mask to obtain inner points from the clothed point cloud (Sec. 3.1); 2) *dense fit*, which encodes the inner points implicitly to establish dense correspondence for SMPL-X model fitting (Sec. 3.2).

Basically, ETCH-X extends ETCH [24] by replacing the SMPL model with the more expressive SMPL-X model [35] and replacing the explicit sparse markers with implicit dense correspondence as in NICP [31], which is not only more robust to partial inputs, but also enables more detailed fitting, particularly for the hands and face. Furthermore, the “undress” and “dense fit” modules are well disentangled and can be trained separately with garment-rich data (*i.e.*, CLOTH3D [3]) and pose-rich data (*i.e.*, AMASS [30] and InterHand2.6M [32]) for robust regression of clothing tightness and body/hand poses, making the training process more flexible and scalable.

3.1 Masked Undress: Clothed to Body Points

Tightness Vector [24]. ETCH proposes to model cloth-to-body tightness using a set of vectors, *i.e.*, tightness vectors. “Tightness vector $\mathbf{v}_i$ ” is the point-wise 3D vector pointing from the outer point $\mathbf{x}_i$ (clothed human body) to the inner point $\mathbf{y}_i$ (underneath body), *i.e.* $\mathbf{y}_i = \mathbf{x}_i + \mathbf{v}_i$. The tightness vector comprises two components: direction $\mathbf{d}_i$ and magnitude b_i , *i.e.* $\mathbf{v}_i = b_i \mathbf{d}_i$. Given a point cloud $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^3\}_N$ that is randomly sampled from the 3D clothed humans, ETCH uses an EPN [11] to produce an $SO(3)$ -equivariant feature $\mathcal{F}$, mean pooling over this feature yields an invariant feature $\overline{\mathcal{F}_{\text{EPN}}}$, or abbreviated as $\overline{\mathcal{F}}$.

The direction highly correlates with human articulated poses, thus it is learned with local approximate SE(3) equivariant features, ensuring the tightness vector consistently maps the cloth surface to the body surface. Specifically, a self-attention network $\mathcal{F}_{\text{self-attn}}$ is used to process the equivariant feature $\mathcal{F}$ over the rotation group dimension (O), to ensure that each group element feature is associated with a group element $\mathbf{g}_j$, then a direction head $\mathcal{F}_{\text{Direc}}$ parameterized with an MLP network followed by transformations is used to process the output feature to help to produce the final direction $\hat{\mathbf{d}}_i$. While the magnitude mainly reflects clothing displacements, which highly correlate with clothing types and body [46, 47], thus, it is learned from the invariant features. From the invariant feature $\overline{\mathcal{F}}$, a *Point Transformer* [58] $\mathcal{F}_{\text{PT-1}}$ is used to capture the contextual information outside each point, then a magnitude MLP head $\mathcal{F}_{\text{Mag}}$ produces $\hat{b}_i$. Finally, the tightness vectors are obtained via $\hat{\mathbf{v}}_i = \hat{b}_i \hat{\mathbf{d}}_i$:

$$\hat{\mathbf{d}}_i = \mathcal{F}_{\text{Direc}}(\mathcal{F}_{\text{self-attn}}(\mathcal{F}_{\text{EPN}}(\mathbf{X})_i)), \hat{b}_i = \mathcal{F}_{\text{Mag}}(\mathcal{F}_{\text{PT-1}}(\overline{\mathcal{F}}(\mathbf{X})_i, \mathbf{x}_i; \delta)). \quad (1)$$

where $\delta = \Theta(\mathbf{x}_i - \mathbf{x}_j)$ is the learned position embedding to encode the relative positions between point pairs $\{\mathbf{x}_i, \mathbf{x}_j\}$.

Tightness Masking. Since not all surface points exhibit non-zero tightness (*i.e.*, regions such as the head, hands, and exposed skin), we introduce a tightness mask for more precise undressing. Determining whether a surface point should exhibit non-zero tightness is naturally a binary classification problem, *i.e.*, we assign a label l_i to each point, ‘1’ for non-zero tightness and ‘0’ for zero tightness. Inspired by ETCH, we use Point Transformer $\mathcal{F}_{\text{PT-2}}$ and $\mathcal{F}_{\text{Label}}$ takes $\overline{\mathcal{F}}(\mathbf{X}) \in \mathbb{R}^{N \times C}$ with position $\mathbf{X} \in \mathbb{R}^{N \times 3}$ as input, and outputs $\mathcal{P} \in \mathbb{R}^{N \times 2}$, represents the probability of a point $\mathbf{x}_i$ belonging to each class (zero vs. non-zero tightness):

$$\mathcal{P}(\mathbf{X}) = \text{softmax}(\mathcal{F}_{\text{Label}}(\mathcal{F}_{\text{PT-2}}(\overline{\mathcal{F}}(\mathbf{X})_i, \mathbf{X}; \delta))), \hat{L} = \text{argmax}(\mathcal{P}(\mathbf{X})). \quad (2)$$

Training and Inference. We regress the direction $\hat{\mathbf{d}}_i$, the magnitude $\hat{b}_i$, and the label $\hat{l}_i$ for each point $\mathbf{x}_i$. The final training loss $\mathcal{L}$ is formulated as follows:

$$\mathcal{L} = w_d \mathcal{L}_d + w_b \mathcal{L}_b + w_l \mathcal{L}_l, \\ \mathcal{L}_d = \sum_{i=1}^N \hat{l}_i \frac{\hat{\mathbf{d}}_i \cdot \mathbf{d}_i}{\|\hat{\mathbf{d}}_i\| \|\mathbf{d}_i\|}, \mathcal{L}_b = \frac{1}{N} \hat{l}_i \sum_{i=1}^N (\hat{b}_i - b_i)^2, \mathcal{L}_l = -\frac{1}{N} \sum_{i=1}^N \log(\mathcal{P}(\mathbf{x}_i, l_i)). \quad (3)$$

During inference, we obtain the inner point $\hat{\mathbf{y}}_i$ by $\hat{\mathbf{y}}_i = \mathbf{x}_i + \hat{l}_i \hat{\mathbf{v}}_i$, where $\hat{\mathbf{v}}_i = \hat{b}_i \hat{\mathbf{d}}_i$ for each point $\mathbf{x}_i$. Finally we get the inner body point clouds $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_i \in \mathbb{R}^3\}_N$ that will be used for the following dense fit stage.

3.2 Dense Fit: Body Points to SMPL-X

SMPL-X [35] extends SMPL with fully articulated hands and an expressive face. As a statistical body model, SMPL-X maps body shape $\beta \in \mathbb{R}^{10}$, facial expression $\psi \in \mathbb{R}^{10}$ and pose $\theta \in \mathbb{R}^{J \times 3}$ parameters to mesh vertices $\mathbf{V} \in \mathbb{R}^{10475 \times 3}$, where

J is the number of human joints ($J = 55$, containing body, eyes, jaw and finger joints in addition to a joint for global rotation). β are linear shape coefficients of the shape blend shape function, and $B_S(\beta)$ accounts for variations of body shapes. θ contains the relative rotation (axis-angle) of each joint plus the root one w.r.t. their parent in the kinematic tree, and $B_P(\theta)$ models the pose-dependent deformation. ψ are PCA coefficients of the expression blend shape function, and $B_E(\psi)$ accounts for variations of facial expressions. Shape displacements $B_S(\beta)$, pose correctives $B_P(\theta)$ and facial expression displacements $B_E(\psi)$ are added together onto the template mesh $\bar{\mathbf{T}} \in \mathbb{R}^{10475 \times 3}$, in the rest pose (or T-pose), to produce the output mesh $\mathbf{T}$:

$$\mathbf{T}(\beta, \theta, \psi) = \bar{\mathbf{T}} + B_S(\beta) + B_P(\theta) + B_E(\psi), \quad (4)$$

Next, the joint regressor $J(\beta)$ is applied to the rest-pose mesh $\mathbf{T}$ to obtain the 3D joints : $\mathbb{R}^{|\beta|} \rightarrow \mathbb{R}^{J \times 3}$. 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 $\mathbf{t} \in \mathbb{R}^3$ as final output $\mathbf{M}$:

$$\mathbf{M}(\beta, \theta, \psi, \mathbf{t}) = W(\mathbf{T}(\beta, \theta, \psi), J(\beta), \theta, \mathbf{t}, \mathcal{W}). \quad (5)$$

Neural ICP (NICP) [31] is a test-time tuning method for human body fitting. Inspired by LVD [15], given an inner point cloud $\mathbf{Y}$, NICP first uses IF-Nets [14] $\mathcal{F}_{\text{IF}}$ to encode it into implicit feature volume $\mathbf{Z}$. Then it learns a neural field $\mathcal{F}_{\text{NF}}$, which is represented by an MLP, given any query point $\mathbf{q} \in \mathbb{R}^3$, the neural field predict the ordered offsets from the query point to the target SMPL-X body model vertices:

$$\mathbf{Z} = \mathcal{F}_{\text{IF}}(\mathbf{Y}), \quad \mathbf{o} = \mathcal{F}_{\text{NF}}(\mathbf{Z}(\mathbf{q})) \quad (6)$$

where $\mathbf{Z}(\mathbf{q})$ denotes the feature queried from the feature volume $\mathbf{Z}$ from position $\mathbf{q}$, and $\mathbf{o} \in \mathbb{R}^N$ denotes the offsets from the position $\mathbf{q}$ to a subset of target SMPL vertices.

Unlike LVD, which predicts offsets to all template vertices (e.g., 10475 for SMPL-X) using a single MLP, NICP introduces LoVD: a local variant with multiple MLP heads, each specialized for a body region (16 regions via spectral clustering). The template vertices are also downsampled by a factor of 10 (e.g., 1051 for SMPL) for efficiency.

NICP preforms test-time fine-tuning for better robustness. Specifically, after obtaining the neural field, NICP fine-tunes it on the input inner point cloud $\mathbf{Y}$ by iterative optimization. First, NICP samples $\mathbf{y}_k$ from $\mathbf{Y}$ as query points, and find its correspondence vertex ID i_k on the SMPL template by

$$i_k = \arg \min_i \|\mathcal{F}_{\text{NF}}(\mathbf{Z}(\mathbf{y}_k))_i\|_2^2 \quad (7)$$

Then it minimizes the distance between correspondence points by updating the parameters θ of the neural field:

$$\theta^* = \arg \min_{\theta} \sum_{k=1}^n \|\mathcal{F}_{\text{NF}}(\mathbf{Z}(\mathbf{y}_k))_{i_k}\|_2^2 \quad (8)$$

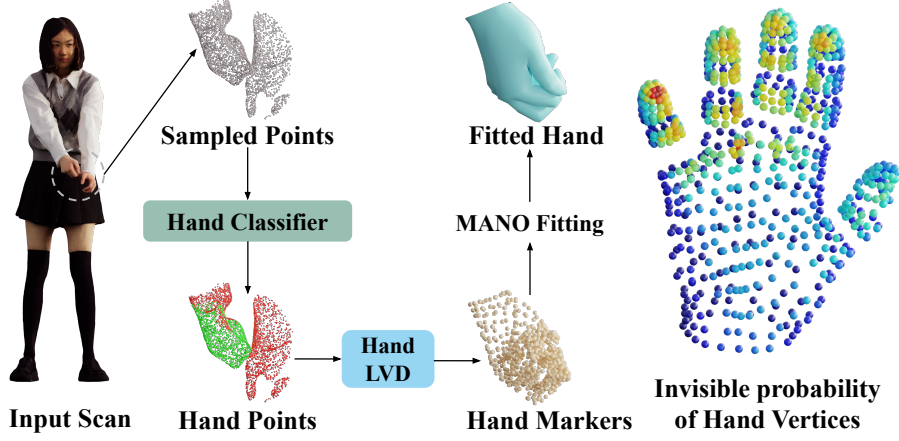


Fig. 3: **Hand Refinement by Re-sampling.** After obtaining initial body fitting, we re-sample points around the hand and fit hand model separately.

The test-time fine-tuning helps improve performance and robustness as depicted in NICP [31]. Then the target SMPL vertices $\{\mathbf{m}_j\}_J$ are obtained by querying the neural field for model fitting. When J vertices are used, the SMPL -X model parameters are fitting by minimizing:

$$\min_{\theta, \beta, \psi, \mathbf{t}} \sum_{j=1}^J \|\mathbf{m}_j - \mathbf{M}(\beta, \theta, \psi, \mathbf{t})_j\|_2^2 \quad (9)$$

Hand Refinement by Re-sampling. Although the hand pose can be fitted while fitting the SMPL-X body, the hand pose is often inaccurate because the sampling points for the hand are usually sparse. To improve hand pose, inspired by some image-based human reconstruction methods that detect and reconstruct hands separately in images [9], we adopted a hand refinement strategy based on re-sampling, as shown in Fig. 3.

Specifically, after obtaining the initial fitted body, we know the approximate position of the hand. Based on this, we resample near the initial hand position to obtain denser hand sampling points. We then individually fit the hand model, MANO [40], to these sampling points. However, as TUCH [33] has shown, the hand easily comes into contact with other body parts, causing the sampled points to include parts that do not belong to the target hand, and these points can affect the hand fit results. Thus, we train a hand classifier on MTP dataset proposed in [33] to remove points that do not belong to the target hand. Then the hand points can be used by a hand LVD network, similar to [15, 31] to produce hand markers, which are finally fitted to MANO hand model. As shown in the right of Fig. 3, due to self-contact and occlusion, many vertices of the hand may be invisible. Therefore, when training the hand LVD, we augment the hand data based on the probability distribution of invisible hand vertices calculated from the MTP dataset. These design strategies are validated in ablation studies Sec. 4.3.

Training and Inference. We follow the same training scheme as NICP [31] to train the LoVD. The neural field predicts 1051 offsets for each query points. At test time, we also perform iterative optimization to fine-tune the LoVD, then obtain the dense correspondence by querying the neural field for following SMPL-X model fitting. After obtaining the fitted SMPL-X body, we perform hand refinement, the fitted MANO hands are then transformed back to the SMPL-X model.

4 Experiments

4.1 Datasets

For fair and comprehensive assessment, we follow established benchmarks [5, 19, 24, 31, 50] and evaluate our method on CAPE [29] and 4D-Dress [51]. Given ETCH-X’s disentangled *undress* and *dense fit* modules, we train them separately on garment-rich (CLOTH3D [3]) and pose-rich (AMASS [30]) datasets to analyze scalability. Our error analysis shows that CAPE and 4D-Dress fitted bodies can be inaccurate, potentially affecting evaluation. To mitigate this, we also benchmark all methods on BEDLAM2.0 [44], which provides simulated clothing with perfect body fits. As no model is trained on BEDLAM2.0, it serves as a pure out-of-distribution (OOD) test, better reflecting generalization. Finally, we describe how we simulate partial scans.

CAPE [29] contains 15 subjects with different body shapes; we split them as 4:1 to evaluate the robustness against *various body shapes and garments*. As NICP [31], we subsample by factors of 5 and 20 for the training and validation sets, resulting in 26,004 train frames and 1,021 valid frames.

4D-Dress [51] has loose clothing and a large range of motion; it contains 32 subjects with 64 outfits across over 520 motions. 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, we obtain 59,395 train frames and 1,943 valid frames.

CLOTH3D [3] is a large-scale simulated dataset of 3D clothed human. It contains a large variability in garment type, topology, shape, size, tightness, and fabric. Dynamic clothes are simulated on top of thousands of different pose sequences and body shapes. It contains more than 2M frames (8K+ sequences) of simulated and rendered garments in 7 categories. We downsampled the full dataset and built roughly 150k paired simulated 3D human scans.

AMASS [30] is a large motion database that unifies different optical marker-based mocap datasets. It contains more than 11,000 motions, covering a wide range of scenarios. As NICP [31], we adopt the official splits and obtain a trainset with roughly 120k SMPL-X bodies by downsampling.

BEDLAM2.0 [44] is a large-scale synthetic video dataset of animated bodies in simulated clothing, containing more than 8M images. it is a significant expansion of the BEDLAM2.0 dataset [44], which increases pose and body BMI variation. It provides complete render assets, including body textures, clothing assets, and

Table 1: **In-distribution Quantitative Comparison with SOTAs.** ETCH-X clearly outperforms SOTAs, whether tightness-agnostic (A.) or -aware (B.), in both CAPE and 4D-Dress across almost all metrics. In 4D-Dress-V2V, it surpasses the ETCH by nearly **21.2%**.

Groups	Methods	CAPE								4D-Dress											
		CD ↓		V2V ↓				MPJPE ↓				CD ↓		V2V ↓				MPJPE ↓			
		All		All	Hands	Head	Other	All	Hands	Head	Other	All		All	Hands	Head	Other	All	Hands	Head	Other
A.	NICP	-		1.736	2.741	1.184	1.827	2.074	2.565	1.042	1.597	-		4.085	6.224	3.323	3.993	4.862	6.142	2.540	3.521
	ArtEq	-		2.202	3.417	2.011	1.943	2.405	3.055	1.693	1.589	-		3.072	4.537	3.145	2.636	3.378	4.170	2.335	2.156
B.	IPNet	1.077		5.529	7.454	5.485	5.001	5.611	6.600	4.527	4.399	1.187		7.495	8.881	7.378	7.178	7.380	8.606	5.973	5.894
	PTF	1.194		2.341	3.880	2.038	2.099	2.641	3.377	1.720	1.767	1.207		3.297	4.938	3.338	2.785	3.567	4.607	2.612	2.248
	ETCH	1.040		1.567	3.449	1.236	1.240	2.002	2.833	0.928	1.007	1.134		2.408	5.108	1.997	2.178	3.459	4.695	1.420	2.141
	ETCH-X	1.015		1.484	2.215	1.120	1.341	1.764	2.148	0.969	1.215	1.060		1.897	3.101	1.836	1.681	2.317	3.065	1.391	1.454

Table 2: **OOD Evaluation.** Note that all methods are trained on 4D-Dress and test on BEDLAM2.0.

Methods	CD ↓	V2V ↓	MPJPE ↓
NICP	-	5.178	6.238
ArtEq	-	4.136	4.447
IPNet	1.369	8.641	9.471
PTF	1.288	3.974	4.668
ETCH	1.454	12.209	15.031
ETCH-X	1.265	3.429	4.033

Table 3: **Evaluation Results for Partial Input.** Notably, single direction chamfer distance (CD) is used here.

Train	Test	CAPE			4D-Dress		
		CD ↓	V2V ↓	MPJPE ↓	CD ↓	V2V ↓	MPJPE ↓
w/o Aug	Full	0.894	1.484	1.764	0.951	1.897	2.317
w/ Aug		0.918	1.644	2.027	0.917	2.135	2.677
Δ		2.7%	10.8%	14.9%	3.6%	12.5%	15.5%
w/o Aug	Partial	1.149	10.056	10.403	2.261	13.861	16.662
w/ Aug		0.951	2.898	3.516	0.978	3.808	5.273
Δ		17.2%	71.2%	66.2%	56.7%	72.5%	68.4%

so on. We randomly sampled 20 subjects with various clothing from the dataset, each with 50 poses, totaling 1,000 paired simulated 3D human scans.

MTF [33] dataset has 3731 images from 148 different subjects, mimicking poses with self-contact sampled from 3DCP Scan, 3DCP Mocap and AGORA. We use MTF data for training hand classifier and computing the probability distribution of invisible hand vertices data augmentation. We also use **InterHand2.6M [32]** dataset to train hand LVD.

Partial Data. We simulate the most common partial data pattern, single-view, *i.e.*, from a certain view angle, only the front part is visible. Given a full mesh, to simulate the single-view mesh, we need to calculate the intersection point of a rays emitted from a specific angle with the mesh surface. We implement this using the Embree [17] library.

4.2 Full-scan Comparison

We compare our method, ETCH-X, with multiple state-of-the-art baselines [5, 24, 31, 50], as shown in Tab. 1, and the qualitative visualization comparison results on 4D-Dress are shown in Fig. 8. Note that ETCH-X predicts SMPL-X bodies while previous methods only predict SMPL bodies, for the fair of comparison, we implement the SMPL-X version of the methods, and all results are calculated based on the SMPL-X body model.

Overall, ETCH-X achieves superior performance across all datasets and metrics. In particular, on CAPE, among all the competitors, our approach reduces the V2V error by **5.3% $\sim$ 73.2%** and MPJPE by **11.9% $\sim$ 68.6%**; on 4D-Dress, the improvement is even more significant with a **21.2% $\sim$ 74.7%** decrease in V2V error and **31.4% $\sim$ 68.6%** in MPJPE. Among tightness-aware methods

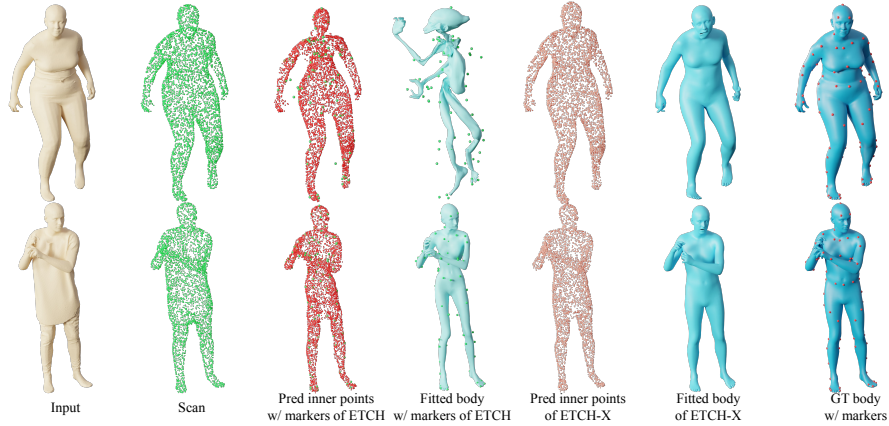


Fig. 4: **Failure Case of ETCH [24] on BEDLAM2.0.** Two representative reasons for ETCH failure are incorrect part labeling (above) and inaccurate inner points (both). The failure is reflected in the large V2V (12.209cm) and MPJPE (15.031cm) errors of ETCH reported in Tab. 2

Table 4: **The Disentangled Design Enables Various Data Sources.**

Methods	Train Data	CAPE			4D-Dress		
		CD ↓	V2V ↓	MPJPE ↓	CD ↓	V2V ↓	MPJPE ↓
NICP	AMASS	-	2.029	2.438	-	5.416	5.280
ETCH	CLOTH3D	1.182	2.465	2.993	1.560	6.989	7.356
ETCH-X	CLOTH3D+AMASS	1.174	1.975	2.365	1.515	4.256	4.200

Table 5: **Ablation Study of Tightness Masking.** The models are trained on CLOTH3D+AMASS.

Methods	CAPE			4D-Dress		
	CD ↓	V2V ↓	MPJPE ↓	CD ↓	V2V ↓	MPJPE ↓
ETCH-X (w/o mask)	1.174	1.975	2.365	1.515	4.256	4.200
ETCH-X (w/ mask)	1.160	1.894	2.266	1.493	4.169	4.083

(*i.e.*, IPNet, PTF, ETCH and Ours), under bidirectional Chamfer Distance, our method achieves 2.4% ~ 16.4% improvement on CAPE and 7.0% ~ 14.3% on 4D-Dress between the predicted inner points/meshes (w/o SMPL-X fitting) and ground-truth SMPL-X bodies.

Beyond in-distribution evaluation, we also assess out-of-distribution (OOD) generalization in Tab. 1 on our BEDLAM2.0 test set, with all methods trained on the 4D-Dress for fairness. ETCH-X demonstrates notably stronger generalization, achieving 1.8% ~ 13.0% lower Chamfer Distance, 13.7% ~ 71.9% lower V2V, and 9.3% ~ 73.2% lower MPJPE. ETCH, in particular, performs poorly on V2V and MPJPE, likely due to limited generalization caused by its entangled architecture design, as illustrated in Fig. 4.

4.3 Ablation Studies

Partial Input. Section 4.1 (Partial Data) details our single-view partial point cloud simulation. During training, we randomly replace 50% of full scans with partial ones. As shown in Tab. 3, partial augmentation improves fitting performance by up to 72.5% on 4D-Dress (V2V metric) for partial inputs, while only slightly reducing accuracy on full scans (maximum 12.5% drop in V2V on 4D-Dress). This highlights the robustness of ETCH-X’s modular design to partial inputs. Qualitative results are shown in Fig. 5.

Disentangled Design. The disentangled design of ETCH-X enables it to effectively integrate simulated garment data and body pose libraries within a unified

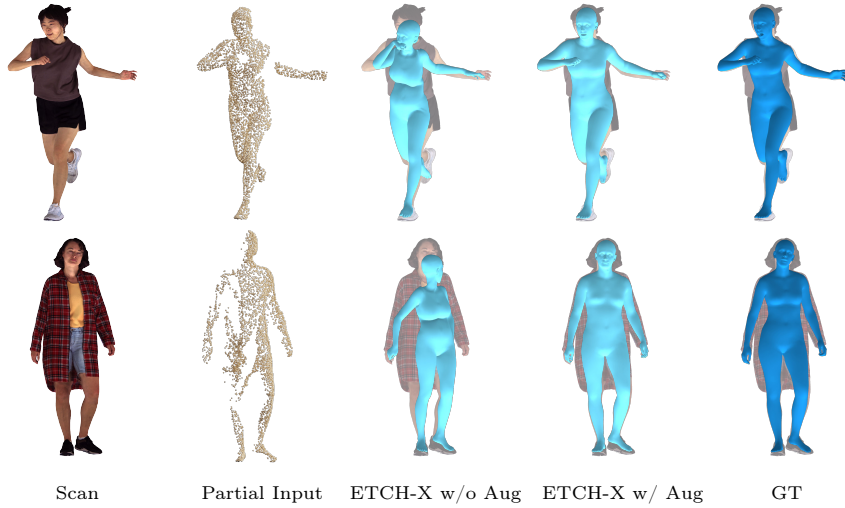


Fig. 5: **Partial Augmentation** . ETCH-X predicts better body poses with partial augmentation.

framework. As demonstrated in Tab. 4, ETCH-X consistently outperforms NICP and ETCH, trained exclusively on either AMASS or CLOTH3D. By leveraging tightness vectors, ETCH-X achieves more accurate undressing, particularly for loose garments in 4D-Dress where NICP often struggles. In contrast, ETCH is limited in pose generalization due to the constrained pose diversity in CLOTH3D.

Scaling Analysis. As discussed in Sec. 1, ETCH-X leverages both simulated garment data (CLOTH3D) and body pose libraries (AMASS), enabling scalability across diverse sources. Figure 6 illustrates performance trends on CAPE and 4D-Dress as the amount of training data increases. For tightness vectors derived from CLOTH3D, performance saturates rapidly on 4D-Dress, whereas adding more AMASS data leads to steady improvements. Scaling holds on CAPE but saturates on 4D-Dress for CLOTH3D likely because its garments are closer in distribution to CAPE (form-fitting) than 4D-Dress (loose). This suggests that the bottleneck is dataset diversity, rather than the method itself. ETCH-X’s performance on loose clothing would likely to improve with more diverse garment simulation data.

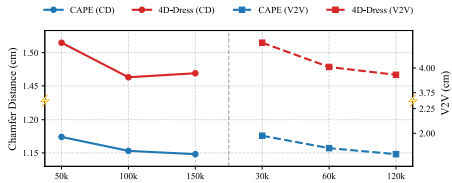
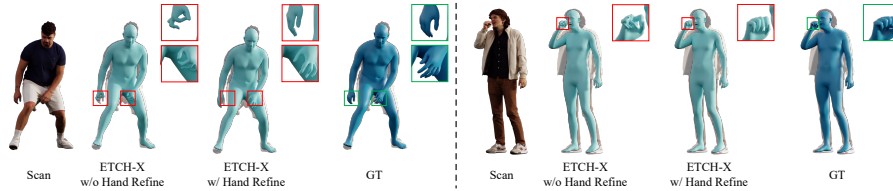


Fig. 6: **Scaling Analysis of ETCH-X**. Increasing the amount of training data from CLOTH3D [3] (left) and AMASS [30] (right) does not necessarily improve performance: tightness accuracy saturates, while pose robustness continues to increase.

Tightness Masking. As described in Sec. 3.1, we introduce a tightness mask to achieve more precise undressing by enforcing zero tightness on exposed skin regions. Since CAPE dataset does not provide body segmentation, we perform tightness masking on CLOTH3D dataset. Quantitative results in Tab. 5 show

Table 6: **Ablation Results of Hand Refinements.**

Settings	Hand LVD	Hand Classifier	Hand Data Augmentation	CAPE				4D-Dress			
				V2V ↓		MPJPE ↓		V2V ↓		MPJPE ↓	
				All	Hands	All	Hands	All	Hands	All	Hands
A.	✗	✗	✗	1.550	2.607	1.948	2.467	1.991	3.417	2.492	3.367
B.	✓	✗	✗	1.514	2.417	1.835	2.277	1.963	3.321	2.444	3.293
C.	✓	✓	✗	1.493	2.278	1.794	2.203	1.928	3.167	2.376	3.125
ETCH-X	✓	✓	✓	1.484	2.215	1.764	2.148	1.897	3.101	2.317	3.065

Fig. 7: **Hand Refinement Results.** ETCH-X produce much better hand poses with hand refinement.

that lower Chamfer Distance (CD) errors indicate inner points that are more body-like, which in turn leads to reduced V2V and MPJPE after fitting.

Hand Refinement. As described in Sec. 3.2, we adopt re-sampling to fit hand separately. The results under different settings are shown in Tab. 6, validating the effectiveness of our design. The visual comparison results in Fig. 3 demonstrates the effect of hand refinement under conditions such as self-contact.

5 Conclusion

We have presented ETCH-X, a novel two-stage pipeline for robustly fitting the SMPL-X body model to clothed 3D scans, regardless of garment type, clothing dynamics, body articulation, or partial observations. By decoupling the process into a masked undress stage and a dense fit stage, our framework remains flexible and scalable with composable synthetic data from diverse sources. However, our approach has limitations, such as efficiency (~ 10 secs for a complete fitting pipeline), and simulated 3D garments have limited diversity. Future work could focus on simulating more diverse 3D garments, and handling more complex scenarios like multi-person interactions and hybrid human-scene LiDAR capture, at real-time speed.

Acknowledgments

We thank all 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.



Fig. 8: Comparison with SOTAs on 4D-Dress.

References

1. Allen, B., Curless, B., Popović, Z.: The Space of Human Body Shapes: Reconstruction and Parameterization from Range Scans. *Transactions on Graphics (TOG)* (2003) [2](#), [5](#)
2. Antić, D., Tiwari, G., Ozcomlekci, B., Marin, R., Pons-Moll, G.: CloSe: A 3D Clothing Segmentation Dataset and Model. In: *International Conference on 3D Vision (3DV)* (2024) [5](#)
3. Bertiche, H., Madadi, M., Escalera, S.: CLOTH3D: Clothed 3D Humans. In: *ECCV* (2020) [4](#), [6](#), [10](#), [13](#)
4. Bhatnagar, B., Petrov, I., Xie, X.: RVH Mesh Registration. https://github.com/bharat-b7/RVH_Mesh_Registration (2022), accessed: 2026-06-28 [5](#)
5. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction. In: *European Conference on Computer Vision (ECCV)* (2020) [2](#), [3](#), [5](#), [10](#), [11](#), [15](#)
6. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Loopreg: Self-supervised Learning of Implicit Surface Correspondences, Pose and Shape for 3D Human Mesh Registration. In: *Conference on Neural Information Processing Systems (NeurIPS)* (2020) [2](#), [3](#), [5](#)
7. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P.V., Romero, J., Black, M.J.: Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In: *European Conference on Computer Vision (ECCV)* (2016) [2](#), [3](#)
8. Cai, Z., Li, Z., Li, X., Li, B., Wang, Z., Zhang, Z., Xiu, Y.: Up2you: Fast reconstruction of yourself from unconstrained photo collections. In: *International Conference on Learning Representations (ICLR)* (2026) [2](#)
9. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Yanjun, W., Pang, H.E., Mei, H., Zhang, M., Zhang, L., Loy, C.C., Yang, L., Liu, Z.: SMPLer-X: Scaling up expressive human pose and shape estimation. In: *Advances in Neural Information Processing Systems* (2023) [9](#)
10. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2019) [3](#), [5](#)
11. Chen, H., Liu, S., Chen, W., Li, H., Hill, R.: Equivariant Point Network for 3D Point Cloud Analysis. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 14514–14523 (2021) [6](#)
12. Chen, X., Pang, A., Yang, W., Wang, P., Xu, L., Yu, J.: TightCap: 3D human shape capture with clothing tightness field. *Transactions on Graphics (TOG)* **41**(1), 1–17 (2021) [5](#)
13. Chen, Y., Medioni, G.: Object Modelling by Registration of Multiple Range Images. *Image and Vision Computing* (1992) [2](#), [5](#)
14. Chibane, J., Alldieck, T., Pons-Moll, G.: Implicit functions in feature space for 3d shape reconstruction and completion. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6970–6981 (2020) [8](#)
15. Corona, E., Pons-Moll, G., Alenyà, G., Moreno-Noguer, F.: Learned Vertex Descent: A New Direction for 3D Human Model Fitting. In: *European Conference on Computer Vision (ECCV)* (2022) [3](#), [8](#), [9](#)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: *International Conference on Learning Representations (ICLR)* (2021) [3](#)

17. Embree: High Performance Ray Tracing Kernels 4.4.0. <https://github.com/RenderKit/embree> (2025), accessed: 2026-06-28 **11**
18. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alphapose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2022) **3, 5**
19. Feng, H., Kulits, P., Liu, S., Black, M.J., Abrevaya, V.F.: Generalizing Neural Human Fitting to Unseen Poses With Articulated SE(3) Equivariance. In: *International Conference on Computer Vision (ICCV)* (2023) **2, 5, 10, 15**
20. de Goes, F., Fong, D., O'Malley, M.: Garment Refitting for Digital Characters. In: *SIGGRAPH Talks* (2020) **2**
21. Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., Lin, L.: Graphonomy: Universal Human Parsing via Graph Transfer Learning. In: *Computer Vision and Pattern Recognition (CVPR)* (2019) **5**
22. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and Inferring Dense Full-Body Human-Scene Contact. In: *Computer Vision and Pattern Recognition (CVPR)* (2022) **2**
23. Jiang, H., Cai, J., Zheng, J.: Skeleton-Aware 3D Human Shape Reconstruction from Point Clouds. In: *International Conference on Computer Vision (ICCV)* (2019) **5**
24. Li, B., Feng, H., Cai, Z., Black, M.J., Xiu, Y.: ETCH: Generalizing Body Fitting to Clothed Humans via Equivariant Tightness. In: *International Conference on Computer Vision (ICCV)* (2025) **2, 3, 4, 5, 6, 10, 11, 12, 15**
25. Li, B., Li, X., Jiang, Y., Xie, T., Gao, F., Wang, H., Yang, Y., Jiang, C.: Garment-dreamer: 3dgs guided garment synthesis with diverse geometry and texture details. In: *International Conference on 3D Vision (3DV)* (2025) **2**
26. Liu, G., Rong, Y., Sheng, L.: 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* (2021) **5**
27. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A Skinned Multi-Person Linear Model. *Transactions on Graphics (TOG)* (2015) **2**
28. Luo, Z., Wang, J., Liu, K., Zhang, H., Tessler, C., Wang, J., Yuan, Y., Cao, J., Lin, Z., Wang, F., et al.: SMPLOlympics: Sports Environments for Physically Simulated Humanoids. *arXiv preprint arXiv:2407.00187* (2024) **2**
29. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to Dress 3D People in Generative Clothing. In: *Computer Vision and Pattern Recognition (CVPR)* (2020) **4, 5, 10**
30. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of Motion Capture As Surface Shapes. In: *International Conference on Computer Vision (ICCV)* (2019) **4, 5, 6, 10, 13**
31. Marin, R., Corona, E., Pons-Moll, G.: NICEP: Neural ICP for 3D Human Registration at Scale. In: *European Conference on Computer Vision (ECCV)* (2024) **2, 3, 4, 5, 6, 8, 9, 10, 11, 15**
32. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: *European Conference on Computer Vision (ECCV)* (2020) **4, 6, 11**
33. Müller, L., Osman, A.A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (Jun 2021) **9, 11**
34. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: Avatars in Geography Optimized for Regression Analysis. In: *Computer Vision and Pattern Recognition (CVPR)* (2021) **5**

35. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In: Computer Vision and Pattern Recognition (CVPR) (2019) [2](#), [3](#), [6](#), [7](#)
36. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: A Model of Dynamic Human Shape in Motion. Transactions on Graphics (TOG) (2015) [2](#), [5](#)
37. Prokudin, S., Lassner, C., Romero, J.: Efficient Learning on Point Clouds with Basis Point Sets. In: International Conference on Computer Vision (ICCV) (2019) [5](#)
38. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In: Computer Vision and Pattern Recognition (CVPR) (2017) [5](#)
39. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In: Conference on Neural Information Processing Systems (NeurIPS) (2017) [5](#)
40. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017) [9](#)
41. Shao, Z., Wang, D., Tian, Q.Y., Yang, Y.D., Meng, H., Cai, Z., Dong, B., Zhang, Y., Zhang, K., Wang, Z.: DEGAS: Detailed Expressions on Full-Body Gaussian Avatars. In: International Conference on 3D Vision (3DV) (2025) [2](#)
42. Shuai, Q., Fang, Q., Dong, J., Peng, S., Huang, D., Bao, H., Zhou, X.: EasyMo-Cap - Make human motion capture easier (2021), <https://github.com/zju3dv/EasyMocap>, accessed: 2026-06-28 [2](#), [3](#), [5](#)
43. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3. arXiv preprint arXiv:2508.10104 (2025) [3](#)
44. Tesch, J., Becherini, G., Achar, P., Yiannakidis, A., Kocabas, M., Patel, P., Black, M.J.: BEDLAM2.0: Synthetic Humans and Cameras in Motion. In: The Thirtieth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025) [4](#), [10](#)
45. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: KPConv: Flexible and Deformable Convolution for Point Clouds. In: International Conference on Computer Vision (ICCV). pp. 6411–6420 (2019) [5](#)
46. Tiwari, L., Bhowmick, B.: Garsim: Particle based neural garment simulator. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2023) [7](#)
47. Tiwari, L., Bhowmick, B., Sinha, S.: Gensim: Unsupervised generic garment simulator. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Deep Learning for Geometric Computing Workshop (2023) [7](#)
48. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual Geometry Grounded Transformer. In: Computer Vision and Pattern Recognition (CVPR) (2025) [3](#)
49. Wang, K., Xie, J., Zhang, G., Liu, L., Yang, J.: Sequential 3D Human Pose and Shape Estimation from Point Clouds. In: Computer Vision and Pattern Recognition (CVPR) (2020) [5](#)
50. Wang, S., Geiger, A., Tang, S.: Locally Aware Piecewise Transformation Fields for 3D Human Mesh Registration. In: Computer Vision and Pattern Recognition (CVPR) (2021) [2](#), [3](#), [5](#), [10](#), [11](#), [15](#)

51. Wang, W., Ho, H.I., Guo, C., Rong, B., Grigorev, A., Song, J., Zarate, J.J., Hilliges, O.: 4D-DRESS: A 4D Dataset of Real-world Human Clothing with Semantic Annotations. In: Computer Vision and Pattern Recognition (CVPR) (2024) [4](#), [5](#), [10](#)
52. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point Transformer V3: Simpler, Faster, Stronger. In: Computer Vision and Pattern Recognition (CVPR) (2024) [5](#)
53. Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H.: Point Transformer V2: Grouped Vector Attention and Partition-based Pooling. Conference on Neural Information Processing Systems (NeurIPS) (2022) [5](#)
54. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In: Computer Vision and Pattern Recognition (CVPR) (2020) [2](#)
55. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In: Computer Vision and Pattern Recognition (CVPR) (2021) [5](#)
56. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep Sets. Conference on Neural Information Processing Systems (NeurIPS) **30** (2017) [5](#)
57. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, Accurate, Human Shape Estimation From Clothed 3D Scan Sequences. In: Computer Vision and Pattern Recognition (CVPR) (2017) [5](#)
58. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point Transformer. In: International Conference on Computer Vision (ICCV) (2021) [5](#), [7](#)
59. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: DeepHuman: 3D Human Reconstruction From a Single Image. In: International Conference on Computer Vision (ICCV) (2019) [5](#)
60. Zhou, B., Franco, J.S., Bogo, F., Tekin, B., Boyer, E.: Reconstructing Human Body Mesh from Point Clouds by Adversarial GP Network. In: Asian Conference on Computer Vision (ACCV) (2020) [5](#)
61. Zuffi, S., Black, M.J.: The Stitched Puppet: A Graphical Model of 3D Human Shape and Pose. In: Computer Vision and Pattern Recognition (CVPR) (2015) [2](#), [5](#)