ETCH: Generalizing Body Fitting to Clothed Humans via Equivariant Tightness

Supplementary Material

Figure R.1. **Cross Dataset Testing (Unseen Garments, Shapes, Poses).** ETCH trained on 4D-Dress can generalize well on unseen THuman2.1 scans and HuGe100K generated 3D humans [74]

R.1. Technical Details

Implementation. We train ETCH using the Adam optimizer with a learning rate of 1e-4 and a batch size of 2. The training process requires 21 epochs on 4D-Dress and 39 epochs on CAPE, taking about 4 days on a single NVIDIA GeForce RTX 4090. The number of sampled points in scans is 5000. We set all loss weights w_d, w_b, w_l, w_c to 1.0. For the EPN network used in Sec. 3.3, the radius is set to 0.4, and the number of layers is 2. The marker-fitting ($K = 86$) process converges in approximately 5 seconds after 80 steps per subject. Following [58], we adopt a two-stage optimization: first optimizing the $\boldsymbol{\beta}[:2]$ and $\boldsymbol{\theta}$ for 30 steps with lr $= 5e-1$, then optimizing all parameters for 50 steps with lr $= 2e-1$.

Anchor Points vs. Scattered Points. It should be noted that all the points are uniformly sampled from the *surface* instead of *vertices*. $\mathbf{Y}$ is sampled from the SMPL mesh $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{t})$. These points $\mathbf{y}_j$ are shot along their normals to intersect the outer surface, termed *Anchor Points* $\mathbf{x}_j$. Then we uniformly sample points $\mathbf{x}_i \in \mathbf{X}$, termed *Scattered Points*. Each scattered point finds its closest anchor point based on geodesic distance $g(\mathbf{x}_i, \mathbf{x}_j; \mathcal{S}_\mathbf{X})$, $\mathcal{S}_\mathbf{X}$ is the outer clothed triangle mesh surface; if this distance is below a threshold (= 0.01), it shares the corresponding inner point $\mathbf{y}_j$ and its corresponding marker $\mathbf{m}_k$. Otherwise, the closest inner point, based on Euclidean distance, is selected instead.

R.2. More Experiments and Discussions

Challenge Subsets. As Sec. 4.1 explains, our train-test split (in Tab. 1) of CAPE is across subjects (unseen shapes and outfits), and 4D-Dress is along the motion sequence (unseen poses). For further evaluation of challenging cases, we test all the models on challenging subsets of 4D-Dress. Specifically, in Tab. R.1, we split three challenging subsets (top 10% outliers) from the test set of 4D-Dress, including *Loose Garments* (thickest outfits measured by tightness magnitude b_i), *Extreme Shapes* (most outlier shapes measured by SMPL shape params $\boldsymbol{\beta}[:3]$), and *Challenging Poses* (most outlier poses measured by SMPL pose params $\boldsymbol{\theta}$). ETCH still leads on ALL these challenging subsets, thus its superior OOD generalization improves the overall performance.

Methods	Loose Garment			Extreme Shape			Challenging Pose		
	V2V ↓	MPJPE ↓	CD ↓	V2V ↓	MPJPE ↓	CD ↓	V2V ↓	MPJPE ↓	CD ↓
Tightness-agnostic									
NICP	9.113	6.940	-	3.906	3.165	-	4.523	3.543	-
ArtEq	3.428	2.746	-	2.137	1.460	-	2.420	1.741	-
Tightness-aware									
IPNet	4.441	2.932	1.652	3.751	3.002	1.240	3.860	2.747	1.273
PTF	3.264	2.341	1.600	3.122	2.645	1.216	2.914	2.221	1.249
Ours	**2.276**	**1.455**	**1.340**	**1.831**	**1.074**	**0.998**	**1.992**	**1.171**	**1.070**

Table R.1. **Challenge Subsets of 4D-Dress.** For further evaluation of challenging cases, we test all the models on challenging subsets of 4D-Dress, including Loose Garment, Extreme Shape and Challenging Pose. See Sec. R.2 for the details on how the challenging subsets of 4D-Dress are filtered. ETCH still leads on ALL these challenging subsets, each corresponding to a specific challenging aspect, thereby its superior OOD generalization improves the overall performance.

Methods	CAPE [39]		4D-Dress [59]	
	V2V ↓	MPJPE ↓	V2V ↓	MPJPE ↓
Tightness-agnostic				
NICP [41]	1.726	1.343	4.754	3.654
NICP† [41]	1.245	1.051	4.738	3.729
Tightness-aware				
Ours	1.647	0.922	**1.939**	**1.116**
Ours†	**1.017**	**0.883**	3.474	2.849

Table R.2. **Chamfer-based Post-refinement.** We adopt the best tightness-agnostic approach, NICP [41], and our ETCH, to further analyze the effectiveness of chamfer-based post-refinement. Notably, † denotes the method w/ chamfer-based post-refinement. The results show that post-refinement *improves* performance on tight clothing (CAPE [39]) but *degrades* it for loose clothing (4D-Dress [59]). Therefore, from application perspective, when clothing styles or fit are uncertain, including the "tightness-vector" and excluding the "post-refinement" will yield plausible results.

Chamfer-based Post-refinement. As we have discussed in Sec. 2 and illustrated in Fig. 2, for minimal clothing scenarios like CAPE [39], aligning with the *outer surface* enhances fitting, but it may harm accuracy for loose clothing, such as 4D-Dress [59], where clothing significantly deviates from the *underlying body*. The results in Tab. R.2 echo our assumption. We ablate the necessity of "Chamfer-based Post-refinement" on NICP [41] and our method, and find that both benefit from such a refinement step on the CAPE dataset, yet our post-refined results still outperform NICP due to better pose initialization. In contrast, for loose clothing 4D-Dress, the significant displacement between body and cloth will dominate the post-refinement, inflating the fitted body (see NICP's results in Fig. 2), ultimately worsening fitting results, as shown in Tab. R.2 (4D-Dress, Ours vs. Ours†).

Shape Accuracy. The body fitting errors arise from pose and shape errors. The joint error MPJPE in Tab. 1 confirmspose accuracy. We evaluate "shape accuracy" in optimization-related

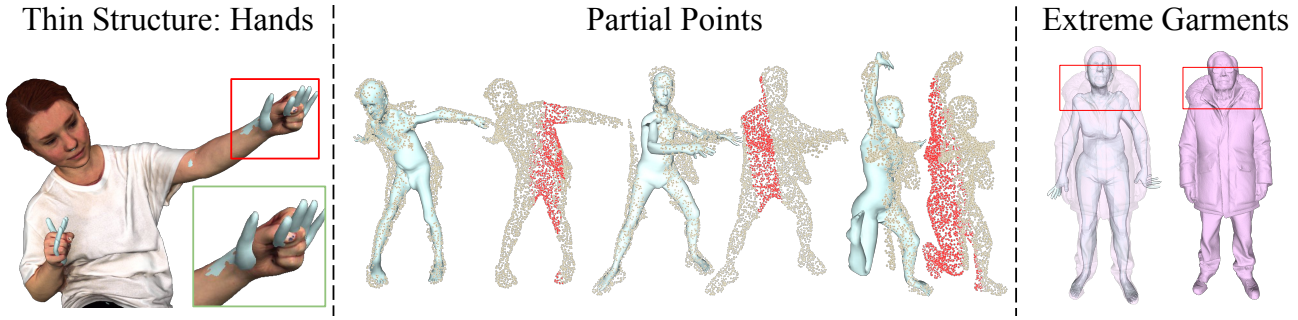

Figure R.2. **Failure Cases.** "Thin Structure" shows that ETCH cannot perfectly fit thin structures such as hands into the scan, since sparse markers do not cover these areas; In "Partial Points", Red points indicate missing partial regions. It leads to incomplete information, which in turn results in failed SMPL fitting; "Extreme Garments" demonstrates a scan featuring a thick fur-trimmed hood, which is rare in the training data but present in the synthetic data. Although ETCH generalizes to different garments, it still struggles with scans that have extremely abnormal appearances or clothing.

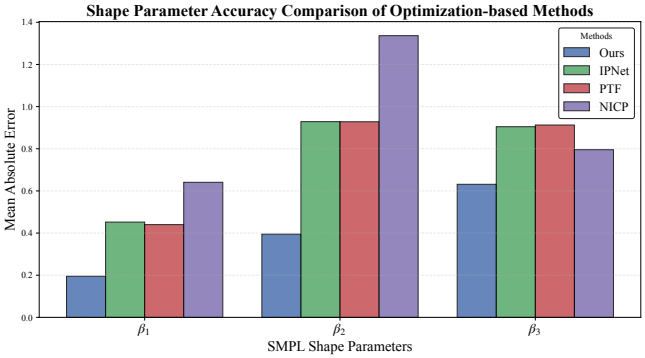

Figure R.3. **Shape Accuracy Analysis.** We calculate the Mean Absolute Error (MAE) on 4D-Dress for the first three principal shape parameters, $\boldsymbol{\beta}_{[:3]} \in \mathbb{R}^3$. Our method shows a significant advantage in shape accuracy, with an average improvement of 49.9%.

methods (*i.e.*, IPNet, PTF, and NICP). ArtEq [22] is excluded for fair comparison as it has access to the ground-truth shape parameters during training. In Fig. R.3, we calculate the Mean Absolute Error (MAE) across 4D-Dress, for the first three principal shape parameters, $\boldsymbol{\beta}_{[:3]} \in \mathbb{R}^3$, which feature key shape information. Our method achieves an average improvement of 49.9%. This performance edge over tightness-aware methods like IPNet and PTF indicates that using tightness vectors to identify the inner body surface, combined with sparse markers, yields more accurate inner body predictions than triple-layer occupancy regression [7], which is also backed by Tab. 1. Even without SMPL fitting step, Chamfer distance (CD) of ETCH is still smaller than PTF and IPNet.

End2End Training. Eq. (8) indicates that our training adopts multitask supervision rather than end-to-end training. Several practical issues prevent us from deploying an end-to-end system: 1) Top-m selection isn't differentiable, and top-all aggregation is time-consuming. 2) End-to-end training may lead to suboptimal shortcuts, affecting useful intermediate outputs for downstream tasks like garment parsing. Thus, multitask supervision remains essential, especially with limited data.

Geodesic-based vs Euclidean-based aggregation We initially considered a Geodesic-based weighted sum but opted for a Euclidean approach due to key limitations: **1) Surface:** The computation of geodesic distance requires a surface, while ETCH uses raw point clouds. **2) Simplicity:** Surface construction (*e.g.* Poisson reconstruction) and geodesic computation are time-consuming. **3) Minor Benefits:** We only select top-3 confidence points for marker aggregation, forming a tiny cluster near the body surface. The Euclidean assumption yields a marker error of just 0.31% relative to body scale. Due to minimal improvement vs. high cost, we chose the Euclidean-based option, which already achieves SOTA performance.

R.3. Limitations and Future Works

While ETCH shows strong performance compared to existing works, it has limitations: **1) Partial Inputs** – As the middle part of Fig. R.2 shows, the reliance on sparse markers means that missing point clouds prevent marker capture and lead to fitting failures. **2) Thin Structures** – The current marker setup does not cover facial landmarks or fingers, and extending it to SMPL-X [37] is non-trivial, as the reception field needs to adapt for full-body, face and hands (refer to the left panel of Fig. R.2) within a unified framework, potentially requiring relaxed tightness modeling when skin is detected. **3) Scalability** – Although we excel in one-shot settings Tab. 3 and achieve state-of-the-art results on CAPE and 4D-Dress, performance gaps narrow with larger data volumes. It remains unclear if performance will plateau at billion-level scan-body pairs, as high-fidelity clothed datasets are scarce and costly. Synthetic [9, 30–32, 66, 75] or body-guided generated humans [34, 35, 63, 74] could be alternatives, but domain gap issues remain.

We will leave all above for future work. In addition, it is straightforward to extend the "tightness vector" to the 2D domain; related topics have already been explored in [18, 25], and Image2Body[†].

[†]https://huggingface.co/spaces/yeq6x/Image2Body_gradio

Figure R.4. **Comparison on Challenging Poses (A, B, C) and Hard Garments (D, E, F, G).** (A) Crossed Legs Pose; (B) Extended Triangle Pose; (C) Asymmetric Limb Pose; (D) Dress Twist; (E) Open Blazer; (F) Flowing Puffer; (G) Waving in Dress. Our method consistently achieves superior pose and shape alignment with ground-truth SMPL. While both ArtEq [22] and ours appear robust to challenging poses – in case A, others misplace the left and right legs; in case B, they misrotate the torso or the head; in case C, the head and the legs are unaligned with ground-truth SMPL – our SMPL results are still better than ArtEq's, particularly in cases A (lower legs), B (raised forearm) and C (left forearm and abdomen). Our advantages are more clear for loose garments. In cases D and F, involving loose clothing and torso rotation, other methods mispredict head or hip rotations, with ArtEq showing a "Taffy and Bowtie" distortion. In case E, they incorrectly predict the left arm position, while in case G, they misplace limbs.