Revisiting Avatar-As-Image: High-Fidelity Registration is All You Need

Anonymous ECCV 2026 Submission

Paper ID #1423

Abstract. Representing 3D clothed humans as standardized 2D UV texture and displacement maps over an underlying body model has long been studied. This compact representation is enticing as it enables pre-trained image networks to process, generate, and edit 3D avatars, but is only useful if scans are accurately aligned and brought into correspondence via high-fidelity registration. This prerequisite has never been met, which we argue explains the limited quality of prior UV-based methods for clothed humans. Despite its significance, no public method produces high-fidelity SMPL(−X)+D registrations with UV texture from arbitrary clothed scans. We present `AvaImg`, a multi-stage optimization pipeline, to close this gap: it enforces body-inside-clothing constraint via signed winding numbers, made viable by a three-level efficiency cascade (∼10× runtime reduced, ∼95% storage saved), and recovers fine surface detail using coarse-to-fine displacement optimization. `AvaImg` outperforms all baselines in body fitting, shape estimation, and surface registration across six datasets, yielding textured registrations near-indistinguishable from scans (PSNR=34.48dB). To substantiate the Avatar-as-Image representation as imminently ready for consume by image foundation models, we auto-encode our UV maps with the frozen VAE from FLUX, achieving only 0.76mm added Chamfer error relative to scan—and indicating that avatar-images lie within the model's distribution, which ultimately conceptually facilitates the use of 2D generative priors for 3D avatar generation. Code, data, and Singularity containers will be publicly released.

Keywords: Digital Human · Surface Registration · Body Fitting

1 Introduction

High-fidelity digital doubles of clothed humans are integral to virtual reality, gaming, telepresence, and generative content creation. Parametric human body models like SMPL(−X) [26, 31] have long been the backbone of digital human modeling. Historically, learning these body and clothing models relied on high-quality registration pipelines, pertaining from SMPL [26] and Dyna [33] to ClothCap [32] and BuFF [42]. Concurrently, 2D generative AI—particularly large-scale diffusion models [8]—has produced powerful visual priors as trained on billions of images. Applying these 2D priors to 3D clothed human generation [37], texture synthesis [21, 35], and editing is now a central goal of the field today.

Fig. 1: Applications enabled by `AvaImg`. Because all registrations share SMPL(−X) parameter space and a coherent UV layout, our output directly supports re-animation, texture editing and transfer, shape editing, dense correspondence, and material editing.

To bridge the gap between 3D geometry and 2D generative models, a natural approach is to represent 3D clothed humans as standard 2D images. Because parametric body models like SMPL(−X) share UV layout across instances, the geometry and appearance of a dressed human can be coherently unrolled onto UV texture and displacement maps. These are standardized images where the same pixel always corresponds to the same anatomical location, non-pertaining to subject. This *Avatar-As-Image* concept is appealing: off-the-shelf image networks and diffusion models can directly process, edit, and generate 3D humans without specialized 3D architectures [3,16,37]. However, previous attempts at UV-based clothed human modeling have yielded poor results, as the prerequisite for quality has never been met: the fidelity of UV maps was limited by the accuracy of the underlying 3D registration. When fitted mesh deviates from scan surface, the unwrapped UV image inherits every geometric and texture artifact; generative models trained on such data cannot learn physically meaningful priors.

Nowadays still, robust public frameworks capable of registering clothed scans to a unified topology with UV texture are largely unavailable. The only publicly available tool is RVH-Mesh-Registration (RMR) [4–6], which supports neither SMPL−X nor UV texture mapping, and uses unsigned distance metrics that cannot help distinguish whether the body approaches the scan from the in- or

outside. More critically, the "ground-truth" registrations shipped with established datasets (BuFF [42], THuman [41], and 2K2K [17]) demonstrate pervasive body-clothing interpenetration: whether the underlying cause is unsigned distances, unreliable surface normals, or other insufficient containment heuristics, the observable result is physically corrupted reference data. Learning-based registration methods [15,24,28] could in principle address this at scale, but they themselves require high-quality registrations for training, constituting a chicken-and-egg problem only optimization-based pipelines can break.

We present `AvaImg`, a multi-stage optimization pipeline which closes this registration gap and thus makes the Avatar-As-Image concept practically viable. Unlike prior pipelines relying on surface normals to resolve body-clothing sidedness [32,42], `AvaImg` enforces body-inside-clothing containment via signed winding numbers [20] that facilitate a volumetric inside/outside test robust to noisy or open scan surfaces—regions where normals often fail. A three-level efficiency cascade (scan decimation, precomputed voxel grids, winding band reduction) reduces runtime by $\sim$10$\times$ and storage by $\sim$95%; enabling applicability at dataset scale. Robust multiview 3D keypoint lifting provides reliable pose initialization even on non-standard configurations, and a coarse-to-fine displacement formulation with fourth-power edge coupling at high resolution recovers sub-centimeter geometric detail which frequently surpasses the fidelity of dataset ground truth. The pipeline handles real-world scan imperfections (noise, missing regions, broken extremities) based on SMPL($-$X) manifold prior and adaptive region-based constraints, and operates reliably across datasets of widely varying scan quality. In consequence to our precise 3D registration, resulting UV maps are adequately clean for modern image generative models: encoding and decoding our UV maps through the frozen VAE of FLUX [8] accumulates only 0.76mm Chamfer error over scan without retraining—which confirms that registration quality is what makes the Avatar-As-Image concept serviceable. Its coherent topology also enables examples of texture transfer, appearance editing, and reanimation, which we demonstrate in Fig. 1. The entire pipeline will be released as a self-contained package via Singularity containers for end-to-end deployment without manual dependency management. In summary, our contributions are:

- `AvaImg`, a multi-stage optimization pipeline for high-fidelity SMPL($-$X)+D registration with UV texture mapping, supporting arbitrary clothed human scans, including noisy and incomplete real-world captures. Code, data, and Singularity containers will be publicly released.
- **Physics-aware body fitting via efficient signed winding numbers**, replacing surface-normal sidedness heuristics of prior work with a volumetric containment constraint, making it practical at dataset scale through a three-level efficiency cascade that reduces runtime by $\sim$10$\times$ and storage by $\sim$95%.
- **A Revisit of Avatar-As-Image representation**, where we show that UV texture and displacement maps produced at `AvaImg`'s registration fidelity successfully encode and decode through a frozen image diffusion VAE, ultimately confirming that high-quality registration is what this long-explored concept has required.

2 Related Work

2.1 Clothed Human Registration

Registration of a parametric body model to clothed 3D scans revolves around a coupled pair of objectives: naked body fitting beneath clothing, and the capture of outer clothing surface in form of per-vertex displacement (SMPL(−X)+D).

Method	Body Fitting	Surface Registration	Texture Registration
Arteq [15]	✔	✘	✘
Etch [24]	✔	✘	✘
IP-Net [5]	✔	✔	✘
LoopReg [6]	✔	✔	✘
PTF [38]	✔	✔	✘
Ours	✔	✔	✔

Table 1: `AvaImg` introduces conjoined capacity for body fitting, surface registration, and texture mapping.

Body Fitting. SMPLify [9] and SMPLify-X [31] both fit body parameters from 2D joint detections via optimization, but are sensitive towards initialization. Learning-based alternatives (LVD [14], ArtEq [15], ETCH [24]) are faster and more robust to pose variation, however, remain *body-centric*: they recover the naked body skeleton and shape without capturing clothing geometry or texture.

Clothed Surface Registration. ClothCap [32] and BuFF [42] register SMPL+D to 4D scan sequences, using surface normals or temporal fusion to resolve body-clothing sidedness, but both require multi-frame input and are not publicly available. RVH-Mesh-Registration (RMR) [4], the only public tool, uses unsigned distances, produces no texture, and is limited to SMPL+H. Learning-based methods (IPNet [5], PTF [38], NICP [28], LoopReg [6]) can produce SMPL+D registrations, yet all require ground-truth registrations for training, creating a chicken-and-egg problem when those registrations themselves contain artifacts. No publicly available pipeline currently yields high-fidelity SMPL(−X)+D registrations with UV texture mapping from arbitrary single-frame clothed scans; `AvaImg` completes this gap.

2.2 2D Representations for 3D Humans

Representation of clothed 3D humans in UV space of parametric body models has been well explored. Early work showed SMPL UV maps can capture both texture and geometry from images or video [1–3, 22, 23, 29], framing shape regression as image-to-image translation in UV space [3] or predicting full 360° textures from partial observations [23]. These same UV maps are presently standard for neural rendering and dynamic character animation [16, 25, 30, 36, 45], where motion-dependent appearance is generated directly in texture space. Recent UV-space methods include SMPLitex [11], Paint-It [21], SCULPT [35] and Chaudhuri *et al*. [13], while encoding 3D geometry as 2D images to leverage diffusion priors is an emerging trend both for general objects [40] and clothed humans [37].

All methods above are limited by registration quality: without high-fidelity SMPL(−X)+D meshes, resulting UV maps inherit geometric artifacts and texture misalignment. `AvaImg` addresses this bottleneck by producing registrations whose UV maps can be imminently consumed by pretrained image VAEs [8] and diffusion models without modification.

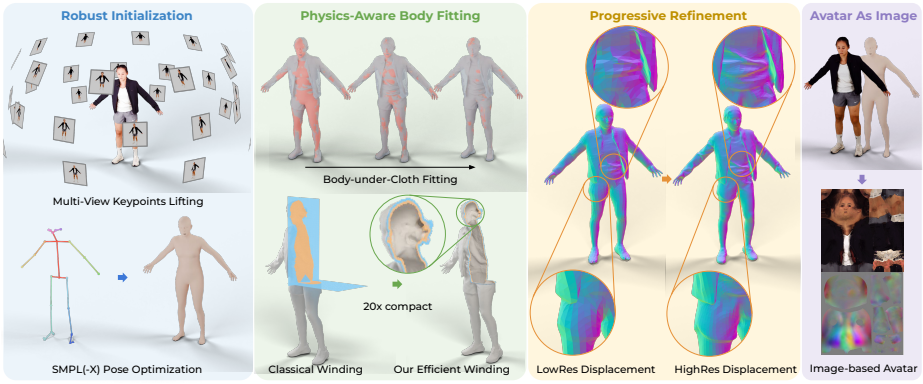

Fig. 2: Method Overview. Given a clothed 3D scan, `AvaImg` proceeds in four stages: (1) multi-view keypoint lifting for robust pose initialization, (2) physics-aware body fitting via efficient signed winding numbers, (3) coarse-to-fine displacement optimization for fine surface detail, and (4) remapping into 2D texture and displacement maps.

3 Method

Given only a raw 3D scan $\mathcal{S} = \{v, f, v_t\}$ with vertices v, faces f, and UV coordinates v_t, `AvaImg` produces a SMPL(−X)+D registration $M(\gamma, \beta, \theta, \psi, D)$ paired with UV texture and displacement maps that jointly define the Avatar-As-Image representation. Three coupled challenges must be resolved for high-fidelity registration: robust pose initialization, naked body fitting below occluding clothing, and complex clothing surface capture.

Two design principles guide every stage of our pipeline. *Coarse-to-fine Optimization*: we constrain heavily at earlier stages to steer away from local minima, then progressively relax constraints to recover fine detail. *Innate Robustness*: we leverage the SMPL(−X) topology itself as a smooth, complete manifold prior, which grounds regularization across scan noise, lacking geometry and merged self-contact regions to produce clean, consistent meshes.

3.1 Robust Pose Initialization

Pose estimation constitutes our pipeline's foundation: errors here onwards propagate irrecoverably across all subsequent stages. Initializing via mean pose, as in prior work [7], routinely traps the optimization in local minima on non-standard configurations, such as raised arms or crossed legs. Instead, we lift reliable 3D joint targets from the scan itself, providing scan-specific initialization.

Scan Normalization. To handle heterogeneous datasets from widely varying coordinate frames, we normalize each scan to gendered SMPL(−X) via a height-based scale factor and centroid translation.

Multi-View Keypoint Lifting. Per textured scan, we render 72 views across 3 elevation levels with full azimuth coverage using PyTorch3D [34]. OpenPose [10] is employed to detect 137 keypoints with respective confidence scores per view. We perform multi-view bundle adjustment to triangulate detected keypoints into

3D joint locations, during which we apply three layers of outlier suppression: the outlier robust $L1$-norm, squaring of the confidence scores, and a hard threshold of 0.3 to zero out unreliable keypoints entirely. To eliminate any inconveniency due to OpenPose runtime compilation, we will provide an end-to-end GPU-friendly Singularity container for user-friendly deployment of our `AvaImg` pipeline.

SMPL(−X) Pose Optimization. The triangulated targets form a joint-fitting loss $\mathcal{L}_j$, which is minimized alongside a shape prior $\mathcal{L}_\beta$ and a pose prior $\mathcal{L}_\theta$ [31]. Following the coarse-to-fine principle, we stage the optimization in three phases: (1) global orientation, head, shoulders, and right foot; (2) all body joints; (3) full SMPL(−X) including hands and facial expression. This schedule yields a robust pose that serves as reliable initialization for body fitting.

3.2 Physics-Aware Fitting with Efficient Winding Numbers

With pose established, we optimize all SMPL(−X) body parameters to situate the naked body beneath clothing; a task complicated by absence of direct observation. Minimizing raw distance between the SMPL(−X) and scan surface commonly finds solution penetrating the clothing outwards as mean to near topology. This is precisely the artifact observed in existing dataset ground truth [17,19,44] and publicly available registration tools.

Inside-Outside Constraint via Signed Winding Numbers. We exploit the physical fact that a body is always enclosed by its clothing. Generalized winding numbers [20] computed on the static scan mesh classify query points as inside (+1) or outside (−1) scan surface. We incorporate this sign into the mesh-to-scan (m2s) distance loss and apply a parametric ReLU:

$$\mathcal{L}_d = \rho(\mathrm{pReLU}[\mathrm{dist}_{\text{signed-m2s}}(\mathcal{S},\, M(\gamma,\beta,\theta,\psi))])\,, \tag{1}$$

where ρ is the Geman-McClure robust function. The pReLU applies asymmetric weighting to the signed distances: body vertices inside of scan clothing receive a mild penalty, while vertices that have escaped outside incur an amplified price, producing a steep, differentiable barrier against penetration. The full objective retains $\mathcal{L}_j$, $\mathcal{L}_\beta$, and $\mathcal{L}_\theta$, with the prior weight annealed over time as to gradually free the body from constraint.

Efficient Winding Computation. Naïve compute of generalized winding numbers is $\mathcal{O}(n_{\text{query}} \times n_{\text{faces}})$, prohibitive on high-polygon scans such as 2K2K [17] (100k–200k faces). We introduce a three-level efficiency cascade that makes the physics constraint practical at dataset scale. *First*, the scan mesh is decimated to approximately 10% of its original face count (min. 40k) for winding computation. *Second*, winding numbers are precomputed on a dense voxel grid (5 mm) around the decimation bounding box and are thresholded to a binary (±1) field. During optimization, classifying a body vertex thus requires only a nearest-neighbor look-up. *Third*, we apply *winding band reduction*: only boundary voxels (those whose 7 nearest neighbors include a sign change) are retained; the remaining $\sim$95% of the voxel grid is discarded, such that both storage and query time are reduced by roughly an order of magnitude.

3.3 Progressive Refinement: From Coarse Body to Fine Detail

Body fitting recovers the underlying body shape but not exterior scan surface: geometric structure, folds, wrinkles, and hair that define visual appearance. We capture these via per-vertex displacements D in a two-pass coarse-to-fine scheme, embodying the progressive refinement principle at mesh resolution level.

Low-Resolution Displacement. With the body now correctly positioned inside the scan and not our focus anymore, the penetration constraint is no longer needed. We switch to unsigned data terms and optimize free vertex positions v_{free} stemming from the SMPL(−X) body mesh, and introduce further loss terms for optimization. $\mathcal{L}_u$ penalizes the difference between displacements in posed and canonical (unposed) space, promoting topology independence from articulation. $\mathcal{L}_c$ is an edge-coupling loss that penalizes edge-length deviations from the non-displaced body, with weighting assigned per select region of bone-based skinning weights of SMPL(−X), restricting deformation in specified areas (*e.g.* hands). $\mathcal{L}_l$ applies cotangent Laplacian smoothing to suppress displacement noise, where weights are assigned by employ of designed vertex maps. *Scheduled weight annealing* allows the mesh to initially capture the global silhouette under stronger regularization, then relaxes to fit finer deforms.

High-Resolution Displacement. The low-resolution displaced mesh is smooth subdivided via Catmull-Clark subdivision [12], approximately quadrupling the face count. Now on high-resolution mesh, a second displacement pass parallel to the one prior recovers fine surface detail via three targeted modifications.

First, the data term switches to unsigned s2m only, directing the mesh toward the outer scan clothing surface without bidirectional pull, where the data adaptive multiplier progressively increases and tightens scan adherence. *Second*, the edge-coupling term $\mathcal{L}_c$ is additionally squared, which produces a loss at a flatter basin but steeper sidewalls: small mesh deviations *snap* towards high-frequency geometry shifts (*e.g.* folds, cloth edges, heels) while maintaining overall consistent topology. *Third*, alterations to region-dependent weights explicitly mitigate where fine structures are perceptually critical and likely otherwise maladaptive. Per experiments in Sec. 4.3, 4.5, we substantiate that our achieved level of geometric fidelity frequently surpasses that of comparable baselines, and can even compensate noise and lack of scan geometric completion via leverage of SMPL(−X) topology as manifold prior (Sec. 4.2)—enabling `AvaImg` to operate reliably across datasets of largely variable scan quality.

3.4 The Avatar-As-Image Representation

The preceding stages yield a high-quality SMPL(−X)+D registration with canonical UV mapping shared across all subjects. To achieve Avatar-As-Image representation, we convert each registration into a pair of standardized images: a UV texture and a UV displacement map per SMPL(−X) UV space, where the same pixel always corresponds to the same anatomical location irrespective of subject. This semantic alignment facilitates the maps as suitable training data for image generative models with no additional alignment or preprocessing.

Precomputed UV Lookup Maps. We precompute an *f-map* and a *b-map* at target image resolution. The f-map specifies which high-res SMPL(−X) UV face each pixel at resolution places within; and the b-map stores the barycentric coordinates of said pixel within respective face. Both maps depend only on the SMPL(−X) UV topology, not on the scan, and are therefore computed once and reused for all subjects. Changing the output resolution (*e.g.* $512^2 \rightarrow 4096^2$) or switching the SMPL(−X) body model variant requires only one-time computation of these lightweight lookup maps.

Texture and Displacement Transfer. For each resolution pixel, the f-map and b-map are used to locate the corresponding 3D point on the high-resolution registered mesh; from which we identify the nearest scan surface point, express it in barycentric coordinates within the closest scan face, and bilinearly sample the scan's color at the corresponding scan UV location. Pixels without coverage are filled in by morphological inpainting. Scans with vertex colors instead of textures interpolate color directly from the nearest scan vertices. Applying the same procedure to per-vertex displacements yields the UV displacement map.

Together, UV texture and UV displacement maps constitute the complete Avatar-As-Image representation. We validate in Sec. 4.4 that they encode and decode through the frozen VAE of a pretrained image diffusion model at high fidelity without any fine-tuning, confirming their readiness as training data for generative models.

4 Experiments

4.1 Evaluation Benchmarks

We evaluate `AvaImg` across six public scan datasets spanning diverse body shapes, poses, and clothing styles: 4D-Dress [39] (47 subjects), BuFF [42] (26), CAPE [27] (40), THuman2.1 [41] (20), 2K2K [17] (20), and CustomHuman [19] (20).

We assess four complementary aspects of registration quality: **Body fitting** is evaluated on the first three datasets via *penetration rate* (% of body vertices outside scan surface), *penetration depth* (mean distance of penetrating vertices to scan surface), and *scan proximity* (mean SMPL(−X) to scan distance) Proximity alone is an insufficient metric, as a body that penetrates the clothing achieves deceptively low distance while being physically implausible. The three body fitting metrics must therefore be interpreted jointly, where low proximity is meaningful only when met by low penetration rate and depth. **Shape estimation** is evaluated on BuFF—which uniquely provides ground-truth minimal-clothing body scans—via bidirectional Chamfer distance in T-pose after Procrustes alignment. **Surface registration** measures the bidirectional Chamfer distance between the registered SMPL(−X)+D mesh and the input scan. **Texture registration** is assessed via multiview rendering PSNR between original scans and textured registrations; we additionally report the Fréchet Inception Distance (FID) [18] as a distributional similarity measure. All quantitative results are summarized in Tab. 3. We validate the Avatar-As-Image concept for latent diffusion model

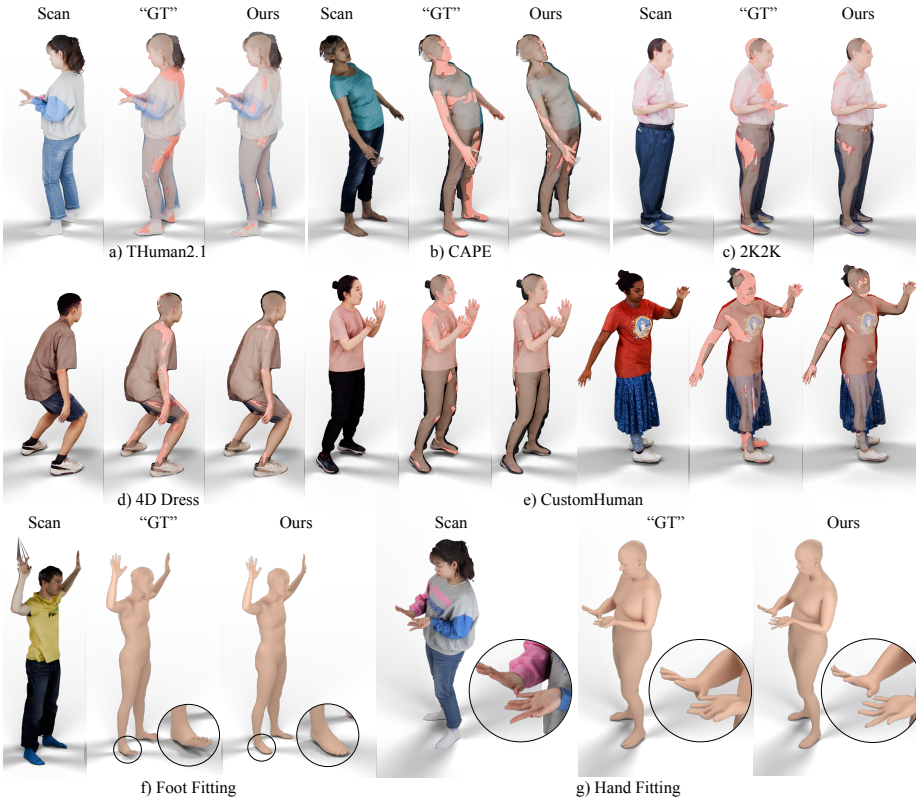

Fig. 3: Comparison with dataset ground truth. Body fitting overlaid on clothed scan surface across five datasets (penetrating vertices highlighted in red). The dataset GT fits exhibit pervasive penetration across the entire body, while ours keep the body enclosed within the clothing (Tab. 2). Zoomed insets show that dataset GT produces bent, misaligned feet and stiff finger articulation, while our method recovers anatomically plausible poses and handles noisy hand geometry.

through a **VAE roundtrip** experiment (Sec. 4.4), and provide **ablation studies** of our key design choices (Sec. 4.5).

4.2 Comparison with Dataset Ground Truth

Several established scan datasets ship "ground-truth" SMPL(−X) registrations alongside their raw scans. We have observed that the provided body-fits exhibit systematic artifacts: fitted body mesh frequently *penetrates* the clothing surface, hand poses are inaccurate due to weak constraints on articulated extremities, and foot alignment is fickle, particularly in CAPE, where open scan surfaces at feet soles provide no geometric anchor (Fig. 3). These are not simply minor cosmetic issues: any downstream method trained on penetrating body-fits obtains a physically impossible prior, and shape estimation benchmarks that evaluate against such registrations risk rewarding methods that replicate the same artifacts.

	Method	2K2K [17]	THuman2.1 [41]	CustomHuman [19]	CAPE [27]	4D-Dress [39]	Overall
Prox. (mm) ↓	Dataset GT	19.5±3.9	**9.2**±1.7	**9.9**±2.0	**6.8**±5.4	12.1±2.7	11.8±4.7
	Ours	**12.9**±2.9	10.0±1.8	11.3±4.4	8.7±1.0	**12.0**±2.9	**11.4**±3.2
Pene. R. (%) ↓	Dataset GT	33.6±9.5	26.2±4.6	33.3±5.7	50.6±7.7	20.9±3.6	28.2±11.3
	Ours	**10.0**±3.2	**15.9**±4.2	**16.1**±6.9	**21.4**±5.6	**14.2**±5.2	**15.0**±6.0
Pene. D. (mm) ↓	Dataset GT	14.5±4.5	5.0±1.0	6.0±1.1	7.4±6.5	4.7±0.8	6.4±4.3
	Ours	**2.1**±0.4	**3.1**±0.4	**4.2**±0.8	**3.6**±0.5	**3.4**±0.4	**3.4**±0.7

Table 2: Body fitting vs. dataset ground truth. Metrics over scan proximity, penetration rate and penetration depth of dataset-provided "ground-truth" SMPL(−X) fits vs. ours, as evaluated on five public datasets. Our body-fits consistently achieve lower penetration metrics, indicating more accurate body-under-clothing estimation.

Improved Body Fitting. By enforcing strict body-inside-clothing containment via signed winding numbers (Sec. 3.2), `AvaImg` eliminates the penetration artifacts present in dataset ground truth. Tab. 2 quantifies this across five datasets: the average penetration rate drops from 28.2% (dataset GT) to **15.0%** (ours), and the average penetration depth decreases from 6.4 mm to **3.4 mm**, while maintaining comparable scan proximity (11.4 mm vs. 11.8 mm). The improvement is most striking on 2K2K, where penetration rate falls from 33.6% to 10.0%, and on CAPE, where it drops from 50.6% to 21.4% despite the challenging open foot surfaces. Our improved fittings can serve as higher-quality training data for learning-based methods [24, 28], breaking the circular dependency (chicken-and-egg problem) identified in Sec. 1.

Robustness to Scan Imperfections. Beyond the quality of the registrations themselves, the input scans are frequently imperfect: *e.g.* BuFF and THuman contain noisy hand geometry, incomplete surface coverage, and missing body regions. Naïve surface fitting propagates these defects into the output mesh, yet `AvaImg` naturally handles such artifacts due to the SMPL(−X) parametric model acting as a strong anatomical prior: the body manifold constrains hands, faces, and extremities to plausible configurations, while the coarse-to-fine displacement scheme (Sec. 3.3) captures underlying surface detail without overfitting to noise. As a result, our registrations are often *cleaner* than the input scans in corrupted regions: the parametric prior effectively denoises the geometry while preserving faithful surface detail elsewhere (Fig. 4).

4.3 Comparison with State-of-the-Art

Body Fitting. We evaluate all methods that produce a naked body fit (ETCH [24], IPNet [5], PTF [38], NICP [28], RVH [7], and ours) on the body fitting metrics defined in Sec. 4.1 across 4D-Dress (47 subjects), BuFF (26) and CAPE (40). Tab. 3 presents the results. RVH and PTF achieve the lowest scan proximity (8.60 mm and 8.68 mm), yet nearly half of their body vertices penetrate the clothing surface (43.3% and 48.9%, respectively). In contrast, our method diminishes the penetration rate to **19.8%**, roughly half that of the next-best method (ETCH, 35.0%), with a penetration depth of only **3.62 mm** at maintain of competitive scan proximity (9.47 mm). This demonstrates that our approach achieves a substantially better trade-off between proximity and physical plausibility: the

	Body Fitting			Shape Est.	Surf. Reg.	Tex. Reg.
Method	Pene. R. (%) ↓	Pene. D. (mm) ↓	Prox. (mm) ↓	Chamfer (mm) ↓	Chamfer (mm) ↓	PSNR ↑
IPNet [5]	44.5	25.93	23.41	9.95	8.61	—
ETCH [24]	<u>35.0</u>	13.53	13.01	<u>7.76</u>	—	—
NICP [28]	54.6	12.98	12.03	8.84	<u>3.06</u>	—
PTF [38]	48.9	8.17	<u>8.68</u>	8.85	6.92	—
RML [7]	43.2	<u>7.80</u>	**8.60**	10.54	3.42	—
Ours	**19.8**	**3.62**	9.47	**7.72**	**2.62**	**34.48**

Table 3: Quantitative comparison. *Body Fitting*: penetration rate, penetration depth, and scan proximity. *Shape Est.*: bidirectional Chamfer distance of T-pose shape to minimum clothing shape in BuFF [42]. *Surf. Reg.*: bidirectional Chamfer distance between registration and scan. *Tex. Reg.*: multiview rendering PSNR between textured registrations and original scans. Best in **bold**, second-best <u>underlined</u>.

estimated body sticks close to the clothing surface while remaining underneath it. Note that optimal proximity is not zero: a correctly estimated naked body should maintain a physical offset from the outer clothing surface. Crucially, our prevailing shape estimation on BuFF **(7.72 mm**, Tab. 3), where ground-truth minimal-clothing body scans are available, confirms that our method does not artificially shrink the body to avoid penetration, but accurately recovers the true underlying body volume.

Surface Registration. We measure clothed surface reconstruction via bidirectional Chamfer distance (100k surface samples per mesh) between SMPL(−X)+D registration and the input scan. We compare IPNet [5], PTF [38], NICP [28], RVH [7], and ours; ETCH is excluded as it produces only naked body parameters. As shown in Tab. 3 (Surf. Reg. column), our method achieves the lowest Chamfer distance across all three datasets, with an overall mean of **2.62 mm**, a 14% reduction over the second-best method (NICP, 3.06 mm). Notably, our method also exhibits the lowest variance (σ=0.39 mm), indicating consistently accurate registrations irrespective of clothing type or body shape. The improvement is most pronounced on 4D-Dress (2.75 mm ours vs. 3.53 mm NICP), which contains the most diverse clothing styles, suggesting that our approach generalizes better to challenging garment geometries. Template-based methods (PTF, IPNet) lag behind, likely because their fixed clothing topology limits the ability to conform to diverse garment shapes.

Texture Registration. Beyond just geometry, our method produces complete textured registration via remapping of scan appearance onto the SMPL(−X) UV layout. To our knowledge, none of the publicly available baselines (IPNet, PTF, NICP, RMR) support texture remapping, making a direct comparison infeasible.

As shown in Tab. 3 (Tex. Reg. column), our method achieves a multiview rendering PSNR of **34.48 dB** against the original scans, confirming high visual fidelity of the textured registrations. In distributional terms, the FID between rendered registrations and rendered scans is only **5.19**, indicating that the two image sets are statistically near-indistinguishable. The low FID confirms that our UV-based texture mapping preserves fine appearance details (*e.g.* fabric pattern, color gradient) with high fidelity, despite the topological transformation from the scan's native mesh to the SMPL(−X) UV parameterization.

Fig. 4: Texture registration and evaluation protocol. Original scan (left) vs. our resulting textured registration (right); insets show clean recovery of noisy hand and head geometry. Both are rendered from multiple viewpoints and compared via image metrics (PSNR = 34.48 dB, Tab. 3).

4.4 Image-based Avatar Representation

Our Avatar-As-Image representation hinges on a key premise: the attained UV maps as produced by `AvaImg` being standardized 2D images which pretrained generative models can both represent and process imminently sans modification. We validate this feasibility by passing our UV texture and displacement maps through the frozen VAE of FLUX [8]—with no fine-tuning—and then measuring reconstruction fidelity in both UV image space (PSNR, SSIM, LPIPS [43]) and 3D geometry space (bidirectional Chamfer distance).

Tab. 4 and Fig. 5 confirm that both image maps survive the roundtrip with minimal degradation. Within 3D space, the VAE adds only **0.76 mm** Chamfer error (3.15 mm → 3.91 mm). UV texture maps achieve a high **38.6 dB** PSNR designating low texture atrophy, and displacement maps reconstruct with **4.98 mm** RMSE, well below the scale of feasibly represented clothing folds as per our supported topology. In spite of the FLUX VAE being trained exclusively on natural images, we recognize only little geometric and visual error.

Thus, these results signify that `AvaImg` UV maps are a ready-to-use input format for image generative architectures, for which encoding and decoding have been shown to be operating faithfully.

	UV Texture		UV Disp.	Spatial	Spatial & Rendering vs. Scan			
	PSNR ↑	SSIM ↑	RMSE (mm) ↓	V2V (mm) ↓	CD (mm) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Before VAE	—	—	—	—	3.15	34.48	0.995	0.006
After VAE	38.6	0.964	4.98	4.29	3.91	30.04	0.988	0.009

Table 4: VAE roundtrip fidelity. UV texture and displacement maps as encoded and decoded through the frozen FLUX VAE [8] vs. scan: bidirectional Chamfer distance and multiview rendering metrics (PSNR, SSIM, LPIPS), both measured against the original scan. The VAE roundtrip adds only 0.76 mm Chamfer error.

Fig. 5: VAE roundtrip validation. (a) Original scan, `AvaImg` registration, and mesh from VAE-decoded UV maps are visually near-indistinguishable. (b) Original UV maps, VAE-reconstructed, and 5× amplified difference.

4.5 Ablation Study

We ablate three key design choices in `AvaImg`: (1) signed vs. unsigned body fitting (Sec. 3.2), (2) coarse-to-fine refinement with fourth-power coupling (Sec. 3.3), and (3) efficient winding computation via decimation and band reduction (Sec. 3.2).

Signed vs. Unsigned Body Fitting. Our full method constrains body vertices to remain *inside* the clothing surface by incorporating the sign of the mesh-to-scan distance via generalized winding numbers [20] (Sec. 3.2). To evaluate the importance of this constraint, we compare against an ablated variant which replaces the signed distance term with a standard unsigned mesh-to-scan distance, identical to the data term used by prior registration pipelines.

Fig. 6 presents the comparison. Without signed winding numbers, the optimizer has no means to distinguish whether the body approaches the scan surface from inside or outside the clothing. As a result, the body mesh frequently *penetrates through* the scan surface to minimize distance, achieving a deceptively low scan proximity at the expense of physical plausibility. With our signed distance formulation, penetration rate drops from 40.7% to **19.8%** and penetration depth from 10.06 mm to **3.62 mm**, while scan proximity remains comparable (9.74 mm vs. **9.47 mm**). The effect extends to shape estimation: on BuFF where ground-truth minimal body shapes are available, shape-under-clothing Chamfer distance drops from 11.95 mm to **7.72 mm**. This confirms that the signed distance constraint is the primary mechanism that prevents body–clothing collision and is essential for physically plausible body estimation.

Progressive Refinement. We ablate the two-pass coarse-to-fine displacement strategy (Sec. 3.3) by comparing three variants on the same 113-subject evaluation set: (i) *low-res only*: displacement optimization of original SMPL(−X)+D mesh without subdivision; (ii) *two-pass, squared coupling*: Catmull-Clark subdivision followed by common squared edge-coupling loss $(w_c\mathcal{L}_c)^2$; and (iii) our full method with fourth-power coupling $(w_c\mathcal{L}_c)^4$ at high resolution. The low-res only variant achieves a surface Chamfer distance of 3.15 mm, as the limited vertex count cannot represent high-frequency clothing detail. Adding subdivision with squared coupling reduces the error to 2.75 mm (−13%), indicating that increased

Distance	Pene. R. ↓	Pene. D. ↓	Prox. ↓	Shape† ↓
Unsigned	40.7%	10.06 mm	9.74 mm	11.95 mm
Signed (ours)	**19.8%**	**3.62** mm	**9.47** mm	**7.72** mm

†BuFF only (26 subj.; GT minimal body required).

Fig. 6: Ablation: signed vs. unsigned body fitting. *Left*: body mesh overlaid on the clothed scan; unsigned distance drives the body through the clothing surface (penetrating vertices in red), while our signed formulation keeps the body enclosed. *Right*: quantitative comparison on 113 subjects (4D-Dress, BuFF, CAPE).

mesh resolution is necessary but not sufficient. Switching to fourth-power coupling (our full method) further lowers the Chamfer to **2.62 mm** (−5%), as the wider penalty basin allows the mesh to conform tighter to fine surface structures instead of smoothing over them. Body fitting and shape metrics are identical across all three variants, as only the clothed surface registration stage differs.

Decimation and Winding Band Reduction. We profiled the two acceleration components of our efficient winding formulation (Sec. 3.2) across decimation levels with and without winding band reduction. Most notably, decimating a select large scan of ∼370k faces down to 40k cuts winding computation time from ∼2,400 s to ∼240 s (∼10×). The winding band reduction at cost of only ∼70 s further accelerates body fitting from ∼1,700 s to ∼240 s (∼7×), and facilitates strong storage saving via decreases of ∼70-100 MB to ∼5 MB (up to 20×). Combined, the physics-aware related stages drop from ∼4,000 s to ∼540 s (∼7×) runtime with no quality degradation.

5 Conclusion

We presented `AvaImg`, a multi-stage optimization pipeline that yields high-fidelity SMPL(−X)+D registrations with UV texture mapping from arbitrary clothed human scans at varying source qualities. By enforcing body-inside-clothing containment via efficient signed winding numbers, our method eliminates the penetration artifacts found in existing dataset ground truth and public registration tools. Evaluations across six datasets confirm that `AvaImg` outperforms all baselines in body fitting, shape estimation, and surface registration. The resulting UV maps encode and decode through a frozen image diffusion VAE with only 0.76 mm added Chamfer error, confirming that registration quality enables the Avatar-As-Image concept.

Limitations and future work. Though slower than feed-forward approaches, as `AvaImg` takes around 30-40 minutes a scan—including 3D joint estimation, winding calculation, body and surface fitting, as well as texture mapping—our high-fidelity registrations can serve as supervision for training fast feed-forward models, while the Avatar-As-Image format itself opens a path toward latent image diffusion models for 3D clothed humans.

References

1. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: 3DV (2018) 4
2. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3D people models. In: CVPR (2018) 4
3. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2Shape: Detailed full human body geometry from a single image. In: ICCV (2019) 2, 4
4. Bhatnagar, B.L.: RVH mesh registration. https://github.com/bharat-b7/RVH_Mesh_Registration (2020), accessed: 2026-03-02 2, 4
5. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3D human reconstruction. In: ECCV (2020) 2, 4, 10, 11
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: NeurIPS (2020) 2, 4
7. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3D people from images. In: ICCV (2019) 5, 10, 11
8. Black Forest Labs: FLUX.1. https://github.com/black-forest-labs/flux (2024) 1, 3, 4, 12
9. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV (2016) 4
10. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE TPAMI **43**(1), 172–186 (2021) 5
11. Casas, D., Comino-Trinidad, M.: SMPLitex: A generative model and dataset for 3D human texture estimation from single image. In: BMVC (2023) 4
12. Catmull, E., Clark, J.: Recursively generated b-spline surfaces on arbitrary topological meshes. Computer-Aided Design **10**(6), 350–355 (1978) 7
13. Chaudhuri, B., Sarafianos, N., Shapiro, L., Tung, T.: Semi-supervised synthesis of high-resolution editable textures for 3D humans. In: CVPR (2021) 4
14. Corona, E., Pons-Moll, G., Alenyà, G., Moreno-Noguer, F.: Learned vertex descent: A new direction for 3D human model fitting. In: ECCV (2022) 4
15. 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: ICCV (2023) 3, 4
16. Habermann, M., Liu, L., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: Real-time deep dynamic characters. ACM TOG **40**(4) (2021) 2, 4
17. Han, S.H., Park, M.G., Yoon, J.H., Kang, J.M., Park, Y.J., Jeon, H.G.: High-fidelity 3D human digitization from single 2K resolution images. In: CVPR (2023) 3, 6, 8, 10
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS. pp. 6626–6637 (2017) 8
19. Ho, H.I., Xue, L., Song, J., Hilliges, O.: Learning locally editable virtual humans. In: CVPR (2023) 6, 8, 10
20. Jacobson, A., Kavan, L., Sorkine-Hornung, O.: Robust inside-outside segmentation using generalized winding numbers. ACM TOG **32**(4) (2013) 3, 6, 13

21. Kim, Y., Oh, T.H., Pons-Moll, G.: Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In: CVPR (2024) 1, 4
22. Lähner, Z., Cremers, D., Tung, T.: DeepWrinkles: Accurate and realistic clothing modeling. In: ECCV (2018) 4
23. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-degree textures of people in clothing from a single image. In: 3DV (2019) 4
24. Li, B., Feng, H., Cai, Z., Black, M.J., Xiu, Y.: ETCH: Generalizing body fitting to clothed humans via equivariant tightness (2025) 3, 4, 10, 11
25. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM TOG **40**(6) (2021) 4
26. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM TOG **34**(6) (2015) 1
27. 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: CVPR. pp. 6468–6477 (2020) 8, 10
28. Marin, R., Corona, E., Pons-Moll, G.: NICP: Neural ICP for 3D human registration at scale. In: ECCV (2024) 3, 4, 10, 11
29. Mir, A., Alldieck, T., Pons-Moll, G.: Learning to transfer texture from clothing images to 3D humans. In: CVPR (2020) 4
30. Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., Habermann, M.: ASH: Animatable gaussian splats for efficient and photoreal human rendering. In: CVPR (2024) 4
31. 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: CVPR (2019) 1, 4, 6
32. Pons-Moll, G., Pujades, S., Hu, S., Black, M.J.: ClothCap: Seamless 4D clothing capture and retargeting. In: ACM TOG. vol. 36, pp. 73:1–73:15 (2017) 1, 3, 4
33. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: A model of dynamic human shape in motion. ACM TOG **34**(4), 120:1–120:14 (2015) 1
34. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3D deep learning with PyTorch3D. In: SIGGRAPH Asia Courses (2020) 5
35. Sanyal, S., Ghosh, P., Yang, J., Black, M.J., Thies, J., Bolkart, T.: SCULPT: Shape-conditioned unpaired learning of pose-dependent clothed and textured human meshes. In: CVPR (2024) 1, 4
36. Sun, G., Dabral, R., Zhu, H., Fua, P., Theobalt, C., Habermann, M.: Real-time free-view human rendering from sparse-view RGB videos using double unprojected textures. In: CVPR (2025) 4
37. Tang, X., Zhang, B., Wonka, P.: Generative human geometry distribution. In: CVPR (2025) 1, 2, 4
38. Wang, S., Geiger, A., Tang, S.: Locally aware piecewise transformation fields for 3D human mesh registration. In: CVPR. pp. 7639–7648 (2021) 4, 10, 11
39. 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: CVPR. pp. 550–560 (2024) 8, 10
40. Yan, X., Lee, H.H., Wan, Z., Chang, A.X.: An object is worth 64x64 pixels: Generating 3D object via image diffusion. In: 3DV (2025) 4

41. 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: CVPR (2021) 3, 8, 10
42. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3D scan sequences. In: CVPR. pp. 4191–4200 (2017) 1, 3, 4, 8, 11
43. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 12
44. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: DeepHuman: 3D human reconstruction from a single image. In: ICCV (2019) 6
45. Zhu, H., Zhan, F., Theobalt, C., Habermann, M.: TriHuman: A real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. ACM TOG **44**(1) (2024) 4