OmniFit: Multi-modal 3D Body Fitting via Scale-agnostic Dense Landmark Prediction

Zeyu Cai[1,5*], Yuliang Xiu[2], Renke Wang[3], Zhijing Shao[4], Xiaoben Li[2], Siyuan Yu[2], Chao Xu[5,6], Yang Liu[5,6], Baigui Sun[5,6], Jian Yang[1], Zhenyu Zhang[1†]

[1]School of Intelligent Science and Technology, Nanjing University
[2]Westlake University [3]NJUST [4]HKUST(GZ)
[5]IROOTECH TECHNOLOGY [6]Wolf 1069 b Lab, Sany Group
Project Page: zcai0612.github.io/OmniFit

Fig. 1: **OmniFit handles diverse 3D human data sources and input modalities. (A)** Raw point clouds from real-world captures. **(B)** Point clouds jointly conditioned on a front-view RGB image. **(C)** Scale-distorted AI-generated 3D human assets. **(D)** Partial point clouds reconstructed from depth maps. For each case, we display the input point cloud alongside the fitted SMPL-X body, demonstrating the robust fitting capability of OmniFit across all modalities and data sources.

Abstract. Fitting an underlying body model to 3D clothed human assets has been extensively studied, yet most approaches focus on either single-modal inputs such as point clouds or multi-view images alone, often requiring known metric scale—a constraint that is frequently unavailable, particularly for AIGC-generated assets where scale distortion is prevalent. We propose OmniFit, where "Omni" signifies our method's ability to seamlessly handle diverse multi-modal inputs (*e.g.*, full scans, partial depth, image captures) while remaining scale-agnostic across both real and synthetic assets. Our key innovation is a simple yet effective conditional transformer decoder that directly transforms surface points into dense body landmarks, which are subsequently used for SMPL-X parameter fitting. Additionally, an optional plug-and-play image adapter enriches geometric details with visual cues to address potential incompleteness. We further introduce a dedicated scale predictor to resize subjects into canonical proportions. Remarkably, OmniFit substantially outperforms

* Work done during an internship at IROOTECH TECHNOLOGY.
† Corresponding author.

state-of-the-art methods by 57.1%–80.9% across daily and loose clothing scenarios, making it the first body fitting method to surpass multi-view optimization baselines and the first to achieve **millimeter-level** accuracy on CAPE and 4D-DRESS benchmarks.

1 Introduction

Fitting an underlying parametric body model [47] to a 3D clothed human asset is a fundamental task in computer vision and graphics, with broad applications in human animation [12, 36–40, 51, 55], avatar creation [11, 23, 32, 52, 60], and AR/VR [9, 14, 59]. Previous body fitting methods have primarily focused on a single input modality, such as RGB images [1, 61], point clouds [6, 62], or depth maps [30]. However, in many real-world scenarios, complementary modalities are simultaneously available—for instance, RGBD cameras that jointly capture color and geometry [24, 63, 69], or AI-based content creation tools that generate textured 3D human meshes [10,15,60,67,74]. These scenarios inspire us to leverage multi-modal sources together for more accurate and robust body fitting.

To this end, we propose OmniFit, a unified framework for fitting SMPL-X body models to 3D clothed human assets across diverse input modalities and data sources. As illustrated in Fig. 1, the name "OmniFit" reflects two key properties: (1) it accepts *multi-modal* inputs, seamlessly handling point clouds alone (A) or jointly with an optional RGB image (B), while gracefully accommodating scale-distorted (C) or partial (D) geometry; (2) it generalizes to *various* clothed human—whether from real-world captures, synthetic datasets, or AI-generated avatars—regardless of pose, body shape, clothing style, or scale.

Unlike traditional optimization-based fitting pipelines that require multi-view rendered images and cascade through multiple error-prone stages (Sec. 2), OmniFit is a 3D-native, learning-based approach built around three key components. At its core is a *landmark predictor* (Sec. 3.2), a conditional transformer decoder that directly estimates a set of predefined dense SMPL-X landmarks from an input point cloud. Inspired by [17], this end-to-end design provides explicit and direct fitting targets for subsequent SMPL-X optimization, making it better suited for data-driven training than previous dense correspondence methods [6, 62]. Moreover, the freely configurable landmark distribution offers comprehensive coverage of fine-grained body regions such as the face and hands, overcoming the limitations of aggregation-based markers [31] that are uniformly distributed and struggle with missing or partial inputs.

To further leverage appearance information available in many 3D human assets, we introduce a *plug-and-play image adapter* (Sec. 3.3) that incorporates a front-view RGB image as an optional additional condition, improving landmark prediction without altering the base architecture. To handle the scale ambiguity that commonly arises in AI-generated assets, we additionally design a *scale predictor* (Sec. 3.4) that estimates a scale factor from a normalized point cloud and restores it to canonical human size prior to landmark prediction.

Our main contributions are:

- We present OmniFit, a unified multi-modal framework for SMPL-X body fitting that seamlessly handles point clouds, optional RGB images, and scale-distorted or partial inputs, generalizing to *various* clothed human across real-world captures, synthetic datasets, and AI-generated assets.
- We introduce three novel modules—a conditional transformer decoder for dense SMPL-X landmark prediction, a plug-and-play image adapter for optional RGB fusion, and a scale predictor for resolving scale ambiguity in AI-generated assets—that together make up OmniFit, enabling accurate, robust, and versatile body fitting.
- To our knowledge, OmniFit is the first 3D body fitting method to achieve **millimeter-level** accuracy, reducing V2V and MPJPE errors by up to 75.5% and 80.9% over the second-best method ETCH [31] on standard benchmarks, while surpassing leading multi-view optimization-based pipelines [1, 61].

2 Related Work

Optimization-based Body Fitting. Early optimization-based methods for body fitting typically leverage the Iterative Closest Point (ICP) algorithm [16] or its variants [2, 48, 76] to align a body template with the target geometry. Modern optimization-based approaches [1, 5, 35, 46, 61, 69, 72, 75] instead operate on rendered multi-view RGB images through a multi-stage pipeline: the input scan is first rendered from multiple viewpoints, followed by 2D keypoint detection (*e.g.*, via OpenPose [13] or AlphaPose [18]); these 2D detections are then lifted to 3D keypoints via multi-view triangulation; finally, SMPL-X parameters are optimized to align with the reconstructed 3D keypoints and point cloud, with optional constraints such as skin segmentation [3, 20] or self-intersection penalties [43]. However, errors in each stage propagate through the pipeline and compound into significant inaccuracies in the final result. Furthermore, these methods fundamentally depend on multi-view rendering, which is inapplicable to untextured meshes or partial point clouds.

Learning-based Body Fitting. Learning-based methods offer efficient alternatives by leveraging large-scale 3D human datasets [4, 35, 41, 63] and deep neural networks [49, 50, 58, 65, 66, 70, 71, 73] tailored for point cloud processing. These approaches either provide initialization for subsequent fitting [6, 7, 62] or directly regress statistical body model parameters [19, 27, 34]. Since direct parameter regression is challenging, most methods rely on intermediate proxy representations, such as joint features [27,34], correspondence maps [7,53,62], part segmentation [19,62], or sparse markers [31]. OmniFit follows the learning-based paradigm and uses predefined dense landmarks as the fitting proxy. Unlike prior methods that take only point clouds as input, OmniFit optionally incorporates a front-view RGB rendering as an additional condition, enabling it to exploit the full information available in 3D human assets.

Correspondence Prediction for Body Fitting. Correspondence prediction is a key step in human body fitting. Explicit dense correspondence methods [6,

7, 42, 62] query pointwise correspondence features from a learned implicit field, establishing a mapping from input points to their corresponding semantic regions on the body template. As dense sampling is computationally expensive, ETCH [31] instead aggregates dense correspondences into a compact set of markers placed at fixed locations on the body template, each representing a cluster center of correspondence features. This design simplifies training but struggles with partial inputs, as missing regions may cause markers to be lost. OmniFit similarly maps input point clouds to predefined landmarks; however, unlike ETCH, landmark locations are not aggregated from dense correspondences. Instead, OmniFit conditions on the input points to directly decode landmark coordinates, with part correspondence implicitly learned and encoded in cross-attention maps.

3 Method

Given a point cloud $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^3\}_{i=1}^N$ of a 3D clothed human, which may be randomly sampled from a mesh, derived from 3DGS point positions, or directly captured by RGBD sensors, our goal is to reconstruct the underlying human body in SMPL-X format (Sec. 3.1). To achieve this, our core solution is fitting SMPL-X parameters to a set of predefined dense landmarks $\mathbf{L} = \{\mathbf{l}_i \in \mathbb{R}^3\}_{i=1}^M$ estimated from the input point cloud $\mathbf{X}$ (Sec. 3.2). To further use the geometric information encoded in the 3D human, we introduce a plug-and-play adapter that enhances landmark prediction with an additional image condition (Sec. 3.3). In addition, to address scale distortion in point cloud—particularly those derived from generative assets—we design a scale prediction module that estimates the scale factor of the input point cloud and rescales it to a canonical human size (Sec. 3.4). Together, these three main modules constitute the OmniFit framework, below we will describe their designs and training details.

3.1 Preliminary

Parametric Body Model – SMPL-X. SMPL-X [47] has been a standard body format in various clothed human datasets [21, 24, 54, 56, 63, 69]. It is a statistical model that maps body shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, pose $\boldsymbol{\theta} \in \mathbb{R}^{J\times 3}$, and facial expression $\boldsymbol{\psi} \in \mathbb{R}^{10}$ 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). $\boldsymbol{\beta}$ are linear shape coefficients of the shape blendshapes, and $B_S(\boldsymbol{\beta})$ accounts for variations of body shapes. $\boldsymbol{\theta}$ contains the relative rotation (axis-angle) of each joint plus the root one w.r.t. their parent in the kinematic tree, and $B_P(\boldsymbol{\theta})$ models the pose-dependent deformation. $\boldsymbol{\psi}$ are PCA coefficients of the expression blend shape function, and $B_E(\boldsymbol{\psi})$ accounts for variations of facial expressions. Shape displacements $B_S(\boldsymbol{\beta})$, pose correctives $B_P(\boldsymbol{\theta})$ and facial expression displacements $B_E(\boldsymbol{\psi})$ 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 $\mathbf{T}$:

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

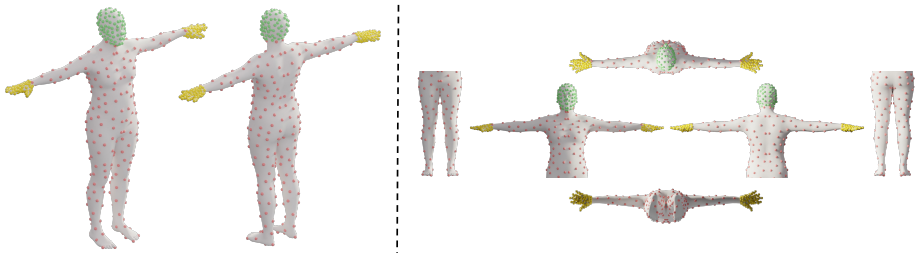

Fig. 2: **Landmark Distribution on the SMPL-X Template.** We place a total of 600 landmarks across three body regions: 120 for the head, 300 for the body, and 180 for the hands.

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

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

3.2 Dense Landmark Prediction

Given a point cloud $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^3\}_{i=1}^N$ of N points, our landmark predictor estimates a set of dense landmarks $\mathbf{L} = \{\mathbf{l}_i \in \mathbb{R}^3\}_{i=1}^M$, each of which corresponds to a predefined vertex on the SMPL-X template mesh. We define $M = 600$ landmarks in total, comprising 120 head, 300 body, and 180 hand landmarks to provide comprehensive coverage of the human body (see Fig. 2).

As shown in Fig. 3, the landmark predictor follows a conditional decoder architecture with two main components: a point cloud encoder and a landmark decoder. The point cloud encoder uses a Point-BERT [70] backbone to tokenize the input into patch embeddings $\mathbf{E} = \{\mathbf{e}_i\}_{i=1}^g$, which are then refined by transformer layers to produce per-patch features $\mathbf{F_P} \in \mathbb{R}^{g\times c}$, where g is the number of patches and c is the feature dimension. The landmark decoder is a Perceiver-style transformer [26] that uses M learnable landmark embeddings $\mathbf{Q} \in \mathbb{R}^{M\times c}$ as queries. Each decoder layer applies:

$$\begin{aligned} i \in [0, l], {\mathbf{F_L}}^0 &= \mathbf{Q}, \\ {\mathbf{F_L}}^i &= \text{CrossAttn}({\mathbf{F_L}}^i, \mathbf{F_P}) + {\mathbf{F_L}}^i, \\ {\mathbf{F_L}}^{i+1} &= \text{SelfAttn}({\mathbf{F_L}}^i) + {\mathbf{F_L}}^i. \end{aligned} \tag{3}$$

CrossAttn$(\cdot)$ and SelfAttn$(\cdot)$ denote the cross-attention and self-attention layers. l indicates the number of transformer layers in the landmark decoder, and ${\mathbf{F_L}}^i \in \mathbb{R}^{M\times c}$ is the hidden states of the i-th layer. After the last layer, a MLP block is applied to the final landmark features ${\mathbf{F_L}}^l$ to obtain the 3D coordinates of the predicted landmarks $\mathbf{L} = \text{MLP}({\mathbf{F_L}}^l) \in \mathbb{R}^{M\times 3}$.

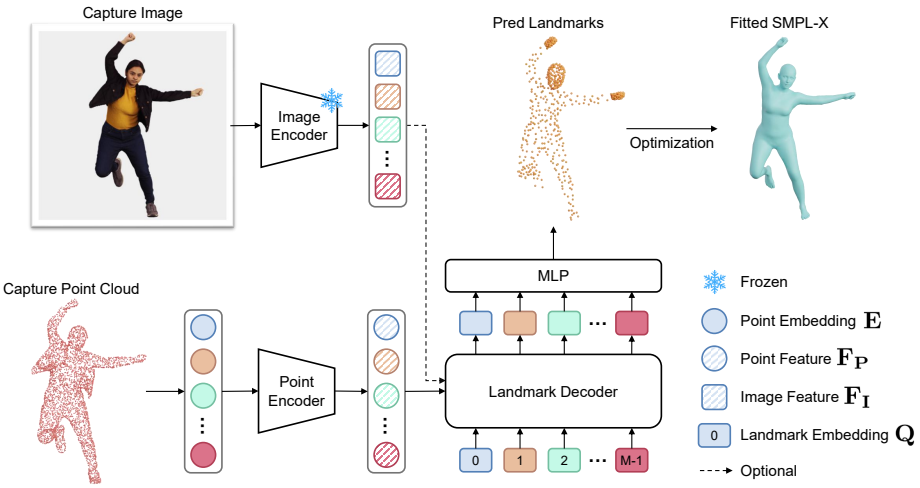

Fig. 3: **The Structure of Landmark Predictor.** Given an input point cloud, the landmark predictor estimates dense 3D landmarks. The point cloud is first tokenized into point embeddings and then fed into the point encoder to produce point features. The landmark decoder is a Perceiver-style transformer that takes learnable landmark embeddings as queries, cross-attends to the point cloud features, and regresses the final 3D landmark coordinates via an MLP. Optionally, image features extracted by a pretrained image encoder can be injected into the landmark decoder, further enhancing the prediction results, which is detailed in Sec. 3.3 and Fig. 4.

In Fig. 3, our landmark predictor can also optionally take an additional condition from extracted image features via a plug-and-play image adapter structure, which is detailed in Sec. 3.3.

3.3 Plug-and-Play Image Adapter

Images are a common companion to 3D human assets, available as rendered RGB images from 3D meshes or 3DGS [28], or RGBD captures of real humans. Compared to point cloud alone, images carry richer details—such as facial features, hand poses, and clothing wrinkles—that benefit accurate body fitting.

Inspired by MV-Adapter [25], we propose a plug-and-play image adapter that takes an optional image $\mathbf{I}$ as an extra input to improve landmark prediction. As shown in Fig. 4, the adapter adds a lightweight cross-attention branch in parallel with the point cloud cross-attention in each decoder layer, with no change to the base architecture. Image features $\mathbf{F_I}$ are extracted by a pretrained image encoder and fused in parallel with the point cloud features. The updated cross-attention in Eq. (3) becomes:

$$\mathbf{F_L}^i = \text{CrossAttn}(\mathbf{F_L}^i, \mathbf{F_P}) + \text{CrossAttn}(\mathbf{F_L}^i, \mathbf{F_I}) + \mathbf{F_L}^i. \tag{4}$$

This design keeps the base model intact, allowing the adapter to be enabled or disabled at will during both training and inference, making the framework flexible for scenarios where images may or may not be available.

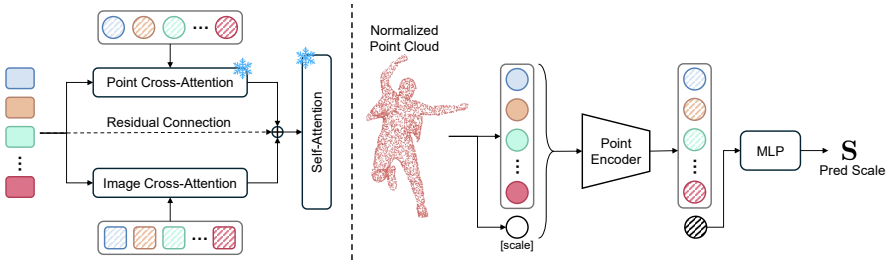

Fig. 4: **Image Adapter (Left) and Scale Predictor (Right). (Left)** The plug-and-play image adapter adds a lightweight cross-attention branch in parallel with the existing point cloud cross-attention in each landmark decoder layer. Image features are fused alongside point cloud features without modifying the base architecture, allowing the adapter to be enabled or disabled at will. **(Right)** The scale predictor shares the same architecture as the point cloud encoder, but adds a scale token to the patch embedding sequence. The token's output is passed through an MLP to regress a scale factor $\mathbf{S}$, which rescales the input point cloud to canonical human size.

3.4 Scale Prediction

3D human assets, particularly those from generative models, often exhibit scale distortion that degrades body fitting accuracy. This arises because generative methods typically normalize outputs to a unit bounding box for training stability, introducing scale ambiguity—for instance, a seated person normalized in the unit box will appear at a larger scale than a canonical standing human.

To address this, we design a scale predictor that takes a normalized point cloud $\tilde{\mathbf{X}}$ of arbitrary pose and estimates a scale factor $\mathbf{S}$ to bring it to canonical human size. As illustrated in Fig. 4, the scale predictor shares a similar architecture with the point cloud encoder, but adds a learnable scale token $\mathbf{E}[\mathbf{s}]$ to the input patch embedding sequence: $\{\mathbf{E}[\mathbf{s}], \tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \ldots, \tilde{\mathbf{e}}_g\}$. After passing through the transformer layers, the output at the scale token position is fed into an MLP to regress $\mathbf{S}$. The rescaled point cloud $\mathbf{X} = \tilde{\mathbf{X}} \cdot \mathbf{S}$ is then passed to the landmark predictor.

3.5 Training and SMPL-X Optimization

The landmark predictor is trained end-to-end using an MSE loss between the predicted landmarks $\mathbf{L} = \{\mathbf{l}_i \in \mathbb{R}^3\}_{i=1}^{M}$ and the ground-truth $\hat{\mathbf{L}} = \{\hat{\mathbf{l}}_i \in \mathbb{R}^3\}_{i=1}^{M}$:

$$\mathcal{L}_{\text{lmk}} = \frac{1}{M} \sum_{i=1}^{M} \|\mathbf{l}_i - \hat{\mathbf{l}}_i\|_2^2. \tag{5}$$

The point cloud encoder and landmark decoder are trained jointly. The image adapter is then attached to the pretrained landmark predictor for adapter training, with the landmark predictor weights frozen and only the adapter branch updated using the same loss in Eq. (5). The scale predictor is trained independently with an MSE loss between the predicted scale factor $\mathbf{S}$ and the ground-truth $\hat{\mathbf{S}}$:

$$\mathcal{L}_{\text{scale}} = \|\mathbf{S} - \hat{\mathbf{S}}\|_2^2. \tag{6}$$

At inference, SMPL-X parameters $\{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{t}\}$ are optimized iteratively by minimizing the landmark alignment objective:

$$\underset{\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{t}}{\operatorname{argmin}} \sum_{i=1}^{M} \|\tilde{\mathbf{l}}_i - \mathbf{l}_i\|_2^2, \tag{7}$$

where $\tilde{\mathbf{l}}_i$ denotes the i-th landmark on the current SMPL-X estimate.

4 Experiments

4.1 Datasets.

Landmark Prediction. For benchmark comparison, we follow the dataset setting and splits in ETCH [31], using the CAPE [35] and 4D-DRESS [63] datasets. For training the generalizable model, we construct a unified dataset by combining CAPE, 4D-DRESS, and two additional synthetic datasets. Specifically, the synthetic data are prepared as follows: (1) We augment the 3D clothed human assets from the BEDLAM2 dataset [57] with hair [22] and shoes, yielding 98,659 frames; (2) We retarget Motion-X [33] pose sequences onto the 3D human templates from the SynBody dataset [68], resulting in 158,410 frames. The inclusion of synthetic data substantially enriches the diversity of human poses and geometric variations, and critically, provides perfectly accurate ground-truth underlying body models, which are essential for training a generalizable model.

Image Adapter. For comparison with multi-view image fitting methods, we use the 4D-DRESS dataset with the same splits as described above. Since 4D-DRESS provides textured 3D human meshes, we render front-view RGB images from these meshes as a condition using orthographic camera projection. We also train a generalizable image adapter with extended training data, including point cloud and front-view rendered image from 2K2K [21] and X-Humans [56].

Scale Prediction. For scale predictor training, we use the unified dataset in landmark prediction, which includes CAPE, 4D-DRESS, and the two synthetic datasets. The input point cloud is normalized to $[-0.9, 0.9]$, and the scale predictor is trained to predict the scale factor that can restore the original size of the input.

Partial Data. Considering the practical scenarios where only partial data may be available, 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 point cloud, we remove the vertices and faces not visible from the front view, and then sample points from the remaining mesh surface.

4.2 Metrics

We evaluate our method using two metrics: (1) **Vertex-to-Vertex (V2V) distance** (cm), which measures the mean Euclidean distance between the predicted

Table 1: **Comparison with Point Cloud Body Fitting Methods.** Our approach outperforms all baselines by a large margin on both CAPE and 4D-DRESS, and is the first method to achieve millimeter-level body fitting accuracy. Compared to previous state-of-the-art ETCH [31], OmniFit reduces V2V error by 57.1%/75.5% and MPJPE by 67.1%/80.9% on CAPE/4D-DRESS. The Δ row reports the per-region relative improvement over the second-based method, with particularly pronounced gains on the hands and head, benefiting from our dense landmark distribution.

Methods	CAPE								4D-DRESS							
	V2V ↓				MPJPE ↓				V2V ↓				MPJPE ↓			
	All	Hands	Head	Body	All	Hands	Head	Body	All	Hands	Head	Body	All	Hands	Head	Body
IPNet	5.529	7.454	5.485	5.001	5.611	6.600	4.527	4.399	7.495	8.881	7.378	7.178	7.380	8.606	5.973	5.894
PTF	2.341	3.880	2.038	2.099	2.641	3.377	1.720	1.767	3.297	4.938	3.338	2.785	3.567	4.607	2.612	2.248
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
ETCH	1.567	3.449	1.236	1.240	2.002	2.833	0.928	1.007	2.408	5.108	1.997	2.178	3.459	4.695	1.420	2.141
Ours	**0.672**	**0.670**	**0.492**	**0.798**	**0.659**	**0.670**	**0.469**	**0.689**	**0.589**	**0.834**	**0.511**	**0.570**	**0.662**	**0.782**	**0.433**	**0.539**
Δ	57.1%	80.6%	60.2%	35.6%	67.1%	76.4%	49.5%	31.6%	75.5%	83.7%	74.4%	78.4%	80.9%	83.3%	69.5%	74.8%

body mesh vertices and the corresponding ground-truth vertices; and (2) **Mean Per Joint Position Error (MPJPE)** (cm), which measures the mean Euclidean error over J SMPL-X body joints. In all reported tables, lower values indicate more accurate body fitting results.

4.3 Implementation Details

The landmark decoder consists of $l = 24$ transformer blocks, each comprising self-attention and cross-attention layers. Both the point cloud encoder and the scale predictor adopt a PointBERT [70] architecture with 16 and 12 transformer blocks, respectively. The image encoder is DINOv2 [44]. All models are trained on the target dataset using the AdamW optimizer with a learning rate of 5×10^{-5} for 100 epochs. The per-GPU batch sizes are set to 20, 24, and 64 for the landmark predictor, image adapter, and scale predictor, respectively. Training is conducted on 8 NVIDIA 48GB GPUs. The number of landmarks is fixed at $M = 600$. During training, the number of input points is randomly sampled from $[5000, 20000]$ with random rotations applied for augmentation, and 50% of the full point clouds are replaced with partial ones to improve robustness to inputs.

Based on [31], we apply a three-stage optimization procedure after landmark prediction: (1) optimizing translation $\boldsymbol{t}$ for 20 steps with $\text{lr} = 5 \times 10^{-1}$; (2) jointly optimizing shape coefficients $\boldsymbol{\beta}[:2]$ and pose coefficients $\boldsymbol{\theta}$ for 30 steps with $\text{lr} = 5 \times 10^{-1}$; and (3) optimizing all parameters for 20 steps with $\text{lr} = 2 \times 10^{-1}$.

4.4 Point Cloud Fitting Comparison

With only point cloud input, we compare out method, OmniFit, with multiple state-of-the-art point cloud body fitting baselines [6, 19, 31, 42, 62], as shown in Tab. 1. Note that OmniFit predicts SMPL-X bodies while previous methods only predict SMPL or SMPL-H bodies. For a fair comparison, we implement the SMPL-X version of these methods, and all the results are evaluated on the SMPL-X body model with the same train and test settings.

Fig. 5: **Qualitative Comparison with Point Cloud Body Fitting Methods.** OmniFit produces more accurate SMPL-X fitting compared to all baselines, especially in the hand and head regions. Failure cases of competing methods are highlighted with red boxes, showing common artifacts such as incorrect hand poses, misaligned head orientations, and inaccurate body shapes. In contrast, OmniFit consistently recovers fine-grained body details across all cases. Please zoom in for a better view.

Overall, OmniFit achieves the best performance across all datasets and metrics, and is the first method to reduce body fitting errors to the millimeter level. In particular, on CAPE, our approach outperforms the second-best method, ETCH [31], by a significant reduction of 57.1% in V2V error and 67.1% in MPJPE; on 4D-DRESS, the improvement is even more significant with a 75.5% reduction in V2V error and 80.9% reduction in MPJPE. These results demonstrate the effectiveness of our conditional landmark decoder design.

The last row of Tab. 1 represents the percentage improvement of our method compared to the second-best method, and it is worth noting that OmniFit has better performance, especially on hands and head regions. We attribute this to the flexible conditional decoder design of the landmark predictor, which allows us to adaptively allocate more landmarks to more complex regions, so the training target will capture finer details and improve final fitting accuracy. Figure 5 shows the visual comparison of the body fitting results, where we can see that OmniFit

Table 2: **Comparison with Multi-View Body Fitting.** OmniFit (point cloud only) already surpasses multi-view methods, and the image adapter further improves accuracy.

	V2V ↓	MPJPE ↓
EasyMocap	3.047	4.029
DiffProxy	2.093	2.630
Ours	0.589	0.662
Adapter	**0.539**	**0.622**

Table 3: **Unified Dataset vs. 4D-DRESS.** Training on large-scale unified data (*) improves generalization to out-of-distribution datasets (CustomHumans, THuman2.1) while maintaining competitive performance on 4D-DRESS.

Methods	4D-DRESS		CustomHumans		THuman2.1	
	V2V↓	MPJPE↓	V2V↓	MPJPE↓	V2V↓	MPJPE↓
Ours	0.589	0.662	1.187	1.280	1.768	1.811
Ours*	0.580	0.643	1.112	1.117	1.338	1.524
Adapter	0.539	0.622	1.181	1.278	1.767	1.806
Adapter*	0.526	0.610	0.878	0.973	0.835	1.144

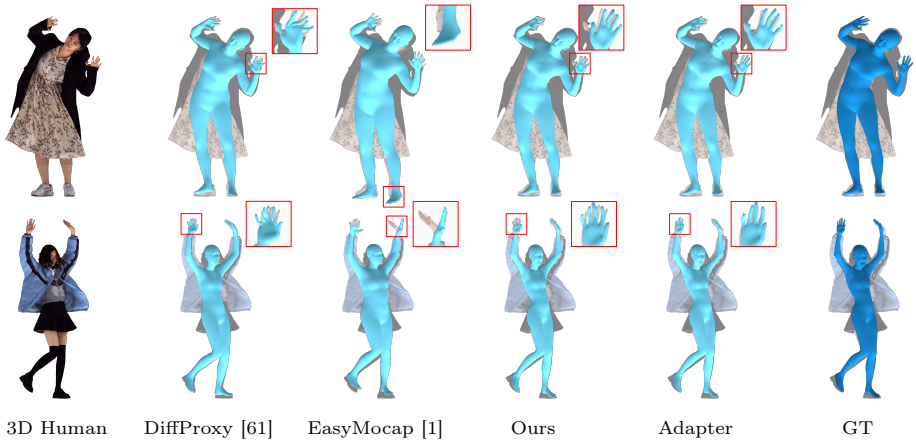

Fig. 6: **Comparison with Multi-View Body Fitting.** OmniFit with the image adapter outperforms multi-view methods DiffProxy [61] and EasyMocap [1] in body fitting accuracy. Details are highlighted with red boxes and magnified, showing that competing methods struggle with fine-grained regions such as hands, while our adapter recovers these details by incorporating point cloud and image.

produces more accurate SMPL-X fitting compared to the baselines, especially in detailed regions like hands and head.

4.5 Image Adapter

The image adapter enables OmniFit to incorporate RGB images as additional input, further improving fitting performance. We evaluate OmniFit with and without the image adapter on the 4D-DRESS dataset, and compare against state-of-the-art multi-view image fitting methods [1, 61], which take multi-view RGB images and camera poses as input. Throughout all tables and figures, **Adapter** denotes the use of the image adapter.

It is worth noting that the comparison is not strictly apples-to-apples. The multi-view baselines rely on rendered RGB observations and camera information, whereas OmniFit operates primarily on mesh-sampled point clouds, optionally augmented with front-view RGB renderings. Our goal is not to claim a direct modality comparison, but to demonstrate that a 3D-native fitting paradigm can

Table 4: **Effectiveness of Scale Predictor.** For scale-normalized 3D humans, incorporating scale predictor improves body fitting accuracy.

	V2V↓	MPJPE↓
gt scale	0.589	0.662
w/o scale pred	1.120	1.172
w/ scale pred	0.651	0.709

Table 5: **Ablation on Landmark Distribution.** Given a total number of 600 landmarks, we evaluate different allocations across body parts and find that setting C. achieves the best overall performance.

	Landmarks			V2V ↓				MPJPE ↓			
Settings	Hands	Head	Body	All	Hands	Head	Body	All	Hands	Head	Body
A.	300	120	180	0.655	**0.832**	0.577	0.657	0.701	0.793	0.499	0.614
B.	240	120	240	0.621	0.833	0.524	0.626	0.690	0.792	0.445	0.597
C.	180	120	300	**0.589**	0.834	**0.511**	**0.570**	**0.662**	**0.782**	**0.433**	**0.539**
D.	120	120	360	0.637	0.955	0.550	0.603	0.724	0.875	0.477	0.558

Fig. 7: **Rescaling and Body Fitting Results on Generated 3D Humans.** For each example, we show (from left to right) the original 3D human, the rescaled 3D human, and the corresponding SMPL-X fitting result. OmniFit also performs well on scale-distorted human assets.

be more effective than render-and-compare optimization when 3D assets are available. Moreover, multi-view images are ultimately projections of 3D geometry, while texture information may not always be available.

As shown in Tab. 2, adding the image adapter consistently improves performance over the point cloud-only backbone and further surpasses leading multi-view image fitting methods. This indicates that point clouds already provide strong geometric cues for body fitting, while image features contribute complementary fine-grained appearance information. The qualitative results in Fig. 6 support this observation, showing more accurate fitting in challenging regions such as the hands. By effectively integrating geometric and visual cues, the image adapter enables OmniFit to achieve superior fitting accuracy compared with methods that rely solely on multi-view images.

4.6 Generalizable Model

To improve the generalization capability of OmniFit, we train a generalizable landmark predictor and image adapter on the unified dataset described in Sec. 4.1. In Tab. 3, we compare the generalizable model with our base model trained only on 4D-DRESS, and evaluate their performance on the test sets of 4D-DRESS and two out-of-distribution datasets, THuman2.1 [69] and CustomHumans [24]. Through training on large-scale datasets, our model achieves better body fitting results for 3D humans of different styles.

4.7 Scale Prediction

In Tab. 4, we demonstrate the effectiveness of the scale predictor. Using the landmark predictor trained on 4D-DRESS for body fitting, we normalize the

Table 6: **Ablation on Number of Landmarks.** More landmarks improve fitting accuracy at the cost of higher computation. We select 600 as a trade-off between accuracy and efficiency.

Num Lmks	V2V↓	MPJPE↓	GFLOPs↓	Iters↑
100	0.637	0.790	**23.82**	**0.88**
300	0.603	0.738	51.65	0.68
600	**0.589**	**0.662**	93.38	0.46

Table 7: **Ablation on Number of Input Points.** Fitting accuracy improves steadily across all the body parts with more input points, demonstrating the scalability of OmniFit to denser point clouds.

Num PCDs	V2V ↓				MPJPE ↓			
	All	Hands	Head	Body	All	Hands	Head	Body
5000	0.589	0.834	0.511	0.570	0.662	0.782	0.433	0.539
10000	0.464	0.676	0.373	0.465	0.531	0.627	0.317	0.442
15000	**0.452**	**0.651**	**0.356**	**0.460**	**0.516**	**0.603**	**0.300**	**0.438**

Table 8: **Ablation on Various Image Encoder Choices.** DINOv2 outperforms Sapiens across all body regions, confirming its effectiveness as the image encoder.

Encoder	V2V ↓				MPJPE ↓			
	All	Hands	Head	Body	All	Hands	Head	Body
Sapiens	0.589	0.836	0.513	0.570	0.661	0.783	0.436	0.533
DINOv2	**0.539**	**0.805**	**0.449**	**0.524**	**0.622**	**0.751**	**0.383**	**0.490**

Table 9: **Results on Partial Inputs.** OmniFit maintains competitive fitting accuracy under partial point cloud.

Test	CAPE		4D-DRESS	
	V2V↓	MPJPE↓	V2V↓	MPJPE↓
Full	0.672	0.659	0.589	0.662
Partial	0.887	1.001	0.595	0.671

input point cloud to $[-0.9, 0.9]$ to simulate the scale setting of 3D generation methods, and subsequently apply the predicted scale factor to restore the original size of the input point cloud. The results show that the scale predictor reduces the V2V error by 41.9% and the MPJPE by 39.5%, substantially outperforming the variant without the scale predictor. Combined with the generalizable landmark predictor, OmniFit is capable of processing diverse 3D humans produced by generative models. As illustrated in Fig. 7, our scale predictor first rescales the generated 3D human to a canonical body size and then performs SMPL-X body fitting, yielding accurate and robust results for 3D humans generated by [15].

4.8 Ablation Studies

We conduct ablation studies to explore the impact of different design choices in OmniFit. For the landmark predictor, we analyze the effect of landmark settings (Tab. 5), number of landmarks (Tab. 6), and number of input points (Tab. 7). For the image adapter, we discuss the choice of image encoder (Tab. 8). We also evaluate the performance of OmniFit with partial inputs (Tab. 9 and Fig. 8) and real-world capture (Tab. 10). We finally visualize the point cross-attention maps in the landmark decoder to verify its part correspondence learning process.

Landmark Setting. As described in Sec. 3.2 and illustrated in Fig. 2, we allocate 600 landmarks in total, distributed as 120 for the head, 300 for the body, and 180 for the hands. To validate this design choice, we compare it against alternative allocation settings. As shown in Tab. 5, our setting achieves the best overall performance, yielding accurate fitting results across all body regions.

Number of Landmarks. Having determined the landmark distribution, we further investigate the effect of the total landmark count. As shown in Tab. 6, increasing the number of landmarks consistently improves fitting accuracy, but

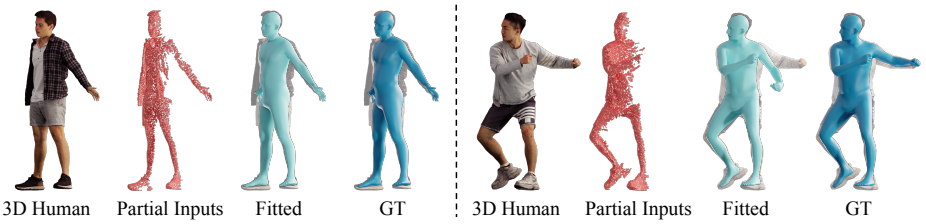

Fig. 8: **Body Fitting Results with Partial Point Cloud Input.** We evaluate OmniFit on partial point cloud inputs. In the right-side case, where entire body regions are missing, OmniFit still produces reasonable SMPL-X fitting results, demonstrating its robustness to incomplete inputs.

Noise Radius	CAPE		4D-DRESS	
(cm)	V2V↓	MPJPE↓	V2V↓	MPJPE↓
2.0	0.838	0.795	0.789	0.926
5.0	1.244	1.158	0.994	1.214

	BEHAVE	
	V2V↓	MPJPE↓
Ours*	2.550	2.838
Adapter*	2.224	2.505

Table 10: **Quantitative Results on Noisy Scans (Left) and BEHAVE Dataset (Right).**

also raises the computational cost (GFLOPs) and reduces training throughput (Iters: training iterations per second). We select 600 landmarks as a favorable trade-off between accuracy and efficiency.

Number of Input Points. For fair comparison with other methods in Tab. 1, we use 5,000 input points uniformly. However, OmniFit is flexible in the number of input points it can handle. As shown in Tab. 7, fitting performance improves steadily with more input points, demonstrating the scalability of our approach.

Choice of Image Encoder. We choose DINOv2 as the image encoder in the plug-and-play image adapter, as it has been widely used in human-related perception tasks [10, 45, 64]. In Tab. 8, we compare DINOv2 with Sapiens [29], which is a recent image encoder designed for human-centric tasks. The results show that DINOv2 performs better than Sapiens, indicating that it is more effective in extracting image features for body fitting.

Partial Inputs. As detailed in Sec. 4.1 (Partial Data), we simulate single-view partial point clouds and incorporate them as supplementary training data alongside full points. Based on this data augmentation, OmniFit retains competitive SMPL-X fitting accuracy under partial inputs, with only an acceptable drop compared to full input (Tab. 9). Figure 8 further provides qualitative results, showing that even when an entire body part is missing, our method still produces reasonable body fitting estimates.

Real-Work Robustness To prove the robustness of OmniFit, we conduct experiments on noisy scans and real-world captures, and the results are shown in Tab. 10. Specifically, we perturb 5K-point inputs with 2K outliers sampled within radii of 2.0 and 5.0 cm, and evaluate the unified-data model and adapter on 20 BEHAVE [8] Kinect captures. OmniFit remains robust under synthetic noise and real-world sensor artifacts, even outperforming the clean-data performance

Fig. 9: **Attention Visualization of Landmarks.** We visualize the point cross-attention maps of the landmark decoder. The attention maps accurately reflect part-level correspondence between input surface points and their associated landmarks (Head, Body, Hand).

of the SOTA method ETCH across datasets (e.g., V2V: 0.994 vs. 2.408 on 4D-DRESS), highlighting its effectiveness on real-world captured data.

Part Correspondence Learning in Attention. Establishing correspondence between surface points and the template body is essential for 3D body fitting [6, 31, 62]. Although our landmark predictor is designed as a conditional decoder that takes the input point cloud as a condition to predict a set of predefined landmarks, its cross-attention maps implicitly capture the correspondence between input surface points and body landmarks. In Fig. 9, we visualize the attention maps of the point cross-attention layers in the landmark predictor, with landmarks grouped by head, body, and hand regions. The results demonstrate that the attention maps accurately reflect part-level correspondence, effectively assigning input points to their respective regions on the template body.

5 Conclusion

We present OmniFit, a unified multi-modal framework for SMPL-X body fitting across diverse types of 3D human assets. OmniFit combines a conditional transformer decoder for dense landmark prediction, a plug-and-play image adapter for optional RGB fusion, and a scale predictor for resolving scale ambiguity in AI-generated assets. It is the first body fitting method to achieve **millimeter-level** accuracy, significantly outperforming prior methods. Despite these advances, OmniFit has two limitations. First, SMPL-X parameters are optimized iteratively from predicted landmarks rather than direct regression, which introduces additional inference time. Second, the image adapter currently supports only a single front-view image as input; extending it to multi-view images could further improve fitting accuracy and robustness. We will leave these as future work to further enhance OmniFit's performance and applicability.

Acknowledgments

This work was supported by the NSFC under Grant No. 62376121, Basic Research Program of Jiangsu under Grant No. BK20251999, Gusu Innovation Leading Talent Program under Grant No. ZXL2025319, and Jiangsu Provincial Science & Technology Major Project under Grant No. BG2024042.

References

1. EasyMoCap - Make human motion capture easier. Github (2021), https://github.com/zju3dv/EasyMocap
2. Allen, B., Curless, B., Popović, Z.: The space of human body shapes: reconstruction and parameterization from range scans. Transactions on Graphics (TOG) (2003)
3. 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)
4. Bertiche, H., Madadi, M., Escalera, S.: Cloth3d: Clothed 3d humans. In: European Conference on Computer Vision (ECCV) (2020)
5. Bhatnagar, B., Petrov, I., Xie, X.: RVH mesh registration. https://github.com/bharat-b7/RVH_Mesh_Registration (2022)
6. 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)
7. 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)
8. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: CVPR (2022)
9. Cai, Z., Huang, Z., Zheng, X., Liu, Y., Liu, C., Wang, Z., Wang, L.: Interact360: Interactive identity-driven text to 360 panorama generation. In: IEEE Conference on Artificial Intelligence (CAI) (2024)
10. 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)
11. Cai, Z., Wang, D., Liang, Y., Shao, Z., Chen, Y.C., Zhan, X., Wang, Z.: Dreammapping: High-fidelity text-to-3d generation via variational distribution mapping. arXiv preprint arXiv:2409.05099 (2024)
12. Cao, Y., Cao, Y.P., Han, K., Shan, Y., Wong, K.Y.K.: Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
13. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2019)
14. Chen, J., Hu, J., Wang, G., Jiang, Z., Zhou, T., Chen, Z., Lv, C.: Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
15. Chen, W., Li, P., Zheng, W., Zhao, C., Li, M., Zhu, Y., Dou, Z., Wang, R., Liu, Y.: SyncHuman: Synchronizing 2d and 3d generative models for single-view human reconstruction. Conference on Neural Information Processing Systems (NeurIPS) (2025)
16. Chen, Y., Medioni, G.: Object modelling by registration of multiple range images. Image and vision computing (1992)
17. Cuevas-Velasquez, H., Yiannakidis, A., Shin, S., Becherini, G., Höschle, M., Tesch, J., Obersat, T., Alexiadis, T., Halilaj, E., Black, M.J.: Mamma: Markerless & automatic multi-person motion action capture. arXiv preprint arXiv:2506.13040 (2025)

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)
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)
20. Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., Lin, L.: Graphonomy: Universal human parsing via graph transfer learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
21. 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
22. He, C., Sun, X., Shu, Z., Luan, F., Pirk, S., Herrera, J.A.A., Michels, D.L., Wang, T.Y., Zhang, M., Rushmeier, H., Zhou, Y.: Perm: A parametric representation for multi-style 3D hair modeling. In: International Conference on Learning Representations (ICLR) (2025)
23. He, X., Wu, Z., Li, X., Kang, D., Zhang, C., Ye, J., Chen, L., Gao, X., Zhang, H., Zhuang, H.: Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement. In: AAAI Conference on Artificial Intelligence (2025)
24. Ho, H.I., Xue, L., Song, J., Hilliges, O.: Learning locally editable virtual humans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
25. Huang, Z., Guo, Y.C., Wang, H., Yi, R., Ma, L., Cao, Y.P., Sheng, L.: Mv-adapter: Multi-view consistent image generation made easy. In: International Conference on Computer Vision (ICCV) (2025)
26. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International Conference on Machine Learning (ICML) (2021)
27. Jiang, H., Cai, J., Zheng, J.: Skeleton-aware 3d human shape reconstruction from point clouds. In: International Conference on Computer Vision (ICCV) (2019)
28. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. Transactions on Graphics (TOG) (2023)
29. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: European Conference on Computer Vision (ECCV) (2024)
30. Lascheit, K., Barath, D., Pollefeys, M., Guibas, L., Engelmann, F.: Robust human registration with body part segmentation on noisy point clouds. arXiv preprint arXiv:2504.03602 (2025)
31. 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)
32. Li, P., Zheng, W., Liu, Y., Yu, T., Li, Y., Qi, X., Chi, X., Xia, S., Cao, Y.P., Xue, W., et al.: Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
33. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. Conference on Neural Information Processing Systems (NeurIPS) (2023)
34. Liu, G., Rong, Y., Sheng, L.: Votehmr: Occlusion-aware voting network for robust 3d human mesh recovery from partial point clouds. In: ACM International Conference on Multimedia (MM) (2021)

35. 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: International Conference on Computer Vision (ICCV) (2020)
36. Ma, Y., Feng, K., Hu, Z., Wang, X., Wang, Y., Zheng, M., He, X., Zhu, C., Liu, H., He, Y., et al.: Controllable video generation: A survey. arXiv preprint arXiv:2507.16869 (2025)
37. Ma, Y., Feng, K., Zhang, X., Liu, H., Zhang, D.J., Xing, J., Zhang, Y., Yang, A., Wang, Z., Chen, Q.: Follow-your-creation: Empowering 4D creation through video inpainting. arXiv preprint arXiv:2506.04590 (2025)
38. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: AAAI Conference on Artificial Intelligence (2024)
39. Ma, Y., Liu, Y., Zhu, Q., Yang, A., Feng, K., Zhang, X., Li, Z., Han, S., Qi, C., Chen, Q.: Follow-Your-Motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. arXiv preprint arXiv:2506.05207 (2025)
40. Ma, Y., Wang, Z., Ren, T., Zheng, M., Liu, H., Guo, J., Fong, M., Xue, Y., Zhao, Z., Schindler, K., et al.: Fastvmt: Eliminating redundancy in video motion transfer. arXiv preprint arXiv:2602.05551 (2026)
41. 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)
42. Marin, R., Corona, E., Pons-Moll, G.: Nicp: Neural icp for 3d human registration at scale. In: European Conference on Computer Vision (ECCV) (2024)
43. Müller, L., Osman, A.A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
44. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal (TMLR) (2024)
45. Örnek, E.P., Labbé, Y., Tekin, B., Ma, L., Keskin, C., Forster, C., Hodan, T.: Foundpose: Unseen object pose estimation with foundation features. In: European Conference on Computer Vision (ECCV) (2024)
46. 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
47. 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
48. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: A model of dynamic human shape in motion. Transactions on Graphics (TOG) (2015)
49. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
50. 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)
51. Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., Qiu, K., Yuan, W., Chen, G., Dong, Z., et al.: Lhm: Large animatable human reconstruction model from a single image in seconds. In: International Conference on Computer Vision (ICCV) (2025)

52. Qiu, L., Zhu, S., Zuo, Q., Gu, X., Dong, Y., Zhang, J., Xu, C., Li, Z., Yuan, W., Bo, L., et al.: Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
53. Sarkar, R., Dave, A., Medioni, G., Biggs, B.: Shape of you: Precise 3d shape estimations for diverse body types. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
54. Shao, R., Pang, Y., Zheng, Z., Sun, J., Liu, Y.: 360-degree human video generation with 4d diffusion transformer. Transactions on Graphics (TOG) (2024)
55. 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)
56. Shen, K., Guo, C., Kaufmann, M., Zarate, J.J., Valentin, J., Song, J., Hilliges, O.: X-avatar: Expressive human avatars. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
57. Tesch, J., Becherini, G., Achar, P., Yiannakidis, A., Kocabas, M., Patel, P., Black, M.J.: BEDLAM2.0: Synthetic humans and cameras in motion. Conference on Neural Information Processing Systems (NeurIPS) (2025)
58. 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) (2019)
59. Wang, B., Meng, H., Cao, R., Cai, Z., Li, L., Ma, Y., Chen, Q., Wang, Z.: MagicScroll: Enhancing immersive storytelling with controllable scroll image generation. In: IEEE Conference Virtual Reality and 3D User Interfaces (VR) (2025)
60. Wang, D., Meng, H., Cai, Z., Shao, Z., Liu, Q., Wang, L., Fan, M., Zhan, X., Wang, Z.: Headevolver: Text to head avatars via expressive and attribute-preserving mesh deformation. In: International Conference on 3D Vision (3DV) (2025)
61. Wang, R., Zhang, Z., Tai, Y., Yang, J.: DiffProxy: Multi-view human mesh recovery via diffusion-generated dense proxies. arXiv preprint arXiv:2601.02267 (2025)
62. Wang, S., Geiger, A., Tang, S.: Locally aware piecewise transformation fields for 3d human mesh registration. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
63. 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
64. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: Prompthmr: Promptable human mesh recovery. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
65. 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
66. 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)
67. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
68. Yang, Z., Cai, Z., Mei, H., Liu, S., Chen, Z., Xiao, W., Wei, Y., Qing, Z., Wei, C., Dai, B., et al.: Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In: International Conference on Computer Vision (ICCV) (2023)

69. 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
70. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
71. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep sets. Conference on Neural Information Processing Systems (NeurIPS) (2017)
72. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
73. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: International Conference on Computer Vision (ICCV) (2021)
74. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. arXiv preprint arXiv:2501.12202 (2025)
75. 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)
76. Zuffi, S., Black, M.J.: The stitched puppet: A graphical model of 3d human shape and pose. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)