DiffProxy: Multi-View Human Mesh Recovery via Diffusion-Generated Dense Proxies

Renke Wang[1], Zhenyu Zhang[*2], Ying Tai[2], Jun Li[*1,2], and Jian Yang[*1,2]

[1] PCA Lab[†], Nanjing University of Science and Technology, Nanjing, China
[2] School of Intelligent Science and Technology, Nanjing University, Nanjing, China

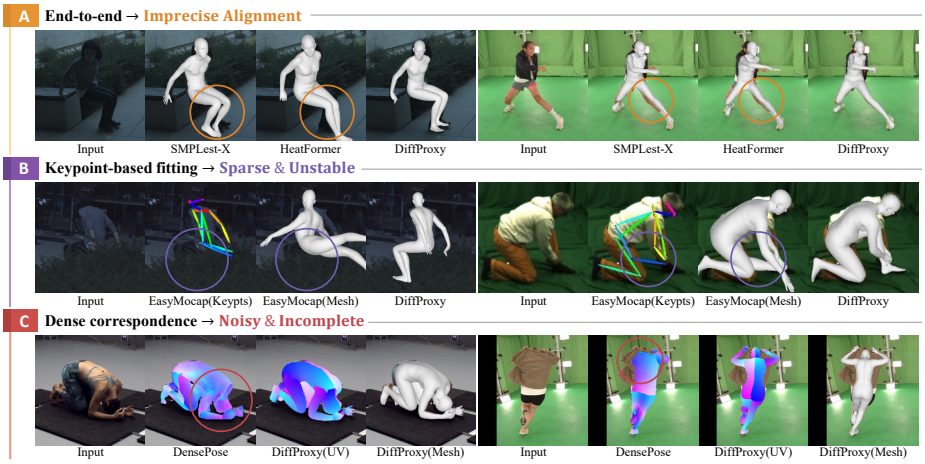

Fig. 1: Existing human mesh recovery methods face distinct challenges. **(A)** End-to-end methods (SMPLest-X, HeatFormer) produce imprecise image-mesh alignment; **(B)** keypoint-based fitting (EasyMoCap) is sparse and unstable under challenging conditions; **(C)** prior dense correspondence predictors (DensePose) produce noisy and incomplete surface mappings. We introduce DiffProxy, which uses a diffusion model trained on synthetic data to predict accurate dense pixel-to-surface proxies, then fits SMPL-X against them via a unified reprojection objective, achieving precise alignment across diverse real-world scenarios.

Abstract. Precise human mesh recovery (HMR) from multi-view images remains challenging: end-to-end methods produce entangled errors hard to localize, while fitting-based methods rely on sparse keypoints that provide limited surface constraints. We observe that the true bottleneck lies in the quality of intermediate representations, and that dense pixel-to-surface correspondences can be effectively generated by repurposing pre-trained diffusion models with rich visual priors. We propose

* Corresponding authors.
† PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Sci. & Tech.

DiffProxy, a Stable-Diffusion-based framework trained on large-scale synthetic data with pixel-perfect annotations. A multi-conditional proxy generator predicts dense correspondences from multi-view images, providing uniform surface constraints that enable precise fitting. Hand refinement feeds enlarged hand crops alongside full-body images for fine-grained detail, while test-time scaling exploits diffusion stochasticity to estimate per-pixel uncertainty. Trained only on synthetic data, DiffProxy achieves state-of-the-art results on five diverse real-world benchmarks. Project page: `https://wrk226.github.io/DiffProxy.html`

1 Introduction

Human mesh recovery (HMR) aims to reconstruct parametric 3D body models from images and has become a cornerstone of computer vision. The dominant paradigm is *end-to-end prediction*: a neural network directly maps an input image to SMPL/SMPL-X [32,40] parameters in a single forward pass [5,11,30,56]. While remarkably fast, this mapping compresses geometry, correspondence, and occlusion reasoning into a single regression target, producing entangled errors that are difficult to localize and often cause image-mesh misalignment [3,6,22].

Not all applications demand real-time speed. In film production, motion capture annotation [1], biomechanical analysis, and digital human creation, precision takes priority. For such scenarios, *fitting-based* methods [1,4,25,40] remain the preferred choice, optimizing body model parameters against detected 2D cues. The canonical multi-view pipeline detects 2D keypoints, triangulates them into 3D, and fits the body model against these 3D keypoints. However, sparse keypoints provide limited constraints: a single erroneous detection or occluded joint can dominate the optimization and produce a completely wrong pose (Fig. 1).

A natural remedy is to replace sparse keypoints with *dense* pixel-to-surface correspondences, which provide far more constraints per view. Since DensePose [13] first established dense image-to-surface mappings, several works [12, 27,51,57] have explored this direction for mesh recovery. Despite the conceptual appeal, these approaches were hindered by two bottlenecks: **(i)** the DensePose-COCO training set relies on manual annotations (∼50K images) that are inherently noisy, propagating label errors into learned models; **(ii)** the CNN-based detectors used in these works lack large-scale visual priors and generalize poorly to out-of-distribution scenarios. As a result, predicted correspondences were insufficiently accurate for high-precision fitting, and the dense-proxy route was largely superseded by end-to-end approaches.

We argue that these bottlenecks reflect limitations of the training data and models available at the time, not inherent flaws of the dense-proxy approach. If we can generate sufficiently accurate dense correspondences, the fitting problem simplifies significantly, as dense pixel-to-surface mappings provide far more constraints per view than sparse keypoints, with uniform coverage across the visible body surface. Recent work has shown that pre-trained diffusion models can be repurposed as powerful dense predictors [24,50], inheriting rich visual priors for cross-domain generalization. This motivates us to revisit the dense-proxy

route by addressing both historical bottlenecks: we train on large-scale synthetic data with pixel-perfect SMPL-X correspondences to eliminate label noise, and leverage Stable Diffusion (SD) 2.1 [45] for synthetic-to-real transfer.

Building on this foundation, we propose **DiffProxy**, a framework for precise multi-view human mesh recovery (Fig. 1). At its core, a *multi-conditional proxy generator* predicts dense pixel-to-surface correspondences from multi-view images, and SMPL-X parameters are recovered by fitting against them via a differentiable reprojection objective. We further introduce *hand refinement*—since hands occupy too few pixels for reliable full-body prediction, we crop and enlarge hand regions as additional views for fine-grained detail via cross-view attention—and *test-time scaling*—for challenging regions such as occlusions or uncommon poses, individual predictions can be erroneous, so we aggregate multiple stochastic samples to derive per-pixel uncertainty for down-weighting unreliable correspondences during fitting. These form a unified pipeline trained entirely on synthetic data.

In summary, our main contributions are:

- We construct a large-scale synthetic multi-view dataset (105K samples, 844K images) with pixel-perfect SMPL-X annotations, eliminating label noise inherent in manual datasets;
- We repurpose a pre-trained diffusion model as a multi-conditional proxy generator for dense pixel-to-surface correspondence prediction, enabling accurate and generalizable proxy estimation from synthetic training alone;
- We introduce hand refinement, which feeds enlarged hand crops as additional views for fine-grained hand recovery, and test-time scaling, which aggregates multiple stochastic predictions to estimate per-pixel uncertainty for robust fitting. Trained exclusively on synthetic data, DiffProxy achieves state-of-the-art results across five diverse real-world benchmarks.

2 Related Work

End-to-end human mesh recovery. Learning-based regression methods directly predict SMPL [32] or SMPL-X [40] parameters from images in a single forward pass. Early CNN-based approaches [22] have evolved into Transformer/ViT architectures [5,11,30,56] trained on large-scale datasets [3,6,29,54]. These methods achieve fast inference but compressing geometry and correspondence reasoning into a single regression target tends to produce entangled errors, limiting achievable precision. Multi-view extensions [21, 28, 35, 59] exploit geometric constraints but often suffer from limited training data and poor cross-dataset generalization. Recent work explores diffusion priors for HMR in parameter [7, 10, 46], mesh [9], or video [15, 58] space. In contrast, we generate dense correspondences as an intermediate representation for precise fitting-based recovery.

Fitting-based human mesh recovery. Optimization-based methods fit parametric body models by minimizing objectives derived from detected cues. SMPLify [4] and SMPLify-X [40] minimize 2D keypoint reprojection errors with pose priors and collision penalties. SPIN [25] combines regression initialization with

iterative fitting. EasyMoCap [1] extends this to multi-view settings with triangulated 3D keypoints. However, these methods depend on sparse keypoints that provide limited surface constraints and are sensitive to detection outliers. Our method replaces sparse keypoints with dense pixel-to-surface correspondences, providing far more constraints per view.

Dense correspondence for mesh recovery. DensePose [13] pioneered pixel-to-surface correspondence for humans, inspiring *iterative fitting* methods such as HoloPose [12] and DenseRaC [51] that optimize SMPL parameters against detected correspondences, and *direct regression* methods like DecoMR [57] and MeshPose [27] that generate meshes in a feed-forward manner. These approaches were hindered by noisy manual annotations in DensePose-COCO ($\sim$50K images) and the poor generalization of deterministic CNN detectors. Our work differs in three key aspects: (i) synthetic training on 105K multi-view samples (844K images in total) with pixel-perfect annotations, eliminating label noise; (ii) diffusion-based generation with strong visual priors for robust generalization; and (iii) multi-view consistency via epipolar attention. Closest in spirit, Look Ma, No Markers [16] is also trained purely on synthetic data and fits a parametric model by minimizing reprojection error across views, but it regresses a fixed set of dense 2D landmarks rather than the per-pixel surface correspondences we use; since its models are not publicly available, we do not compare directly.

Diffusion models for dense prediction. Pre-trained diffusion models have been successfully repurposed for dense prediction tasks. Marigold [24] and GenPercept [50] demonstrated that Stable Diffusion [45] backbones can be adapted for monocular depth and surface normal estimation while retaining zero-shot generalization. In parallel, enforcing multi-view consistency in diffusion models has been explored [18–20, 23, 31, 43, 52] for novel view synthesis and 3D generation, with SPAD [23] introducing epipolar-constrained attention for cross-view interaction. We apply these principles to human dense correspondence prediction, leveraging diffusion priors to generate pixel-to-surface proxies for fitting.

Synthetic data for human mesh recovery. Large-scale synthetic datasets [3, 39, 53, 55] with precise SMPL-X annotations can approach or match real-data performance. We follow this paradigm with multi-view rendering, richer scene diversity (realistic occlusions, varied lighting, and clothing simulation), and pixel-perfect dense correspondence annotations, demonstrating zero-shot generalization to five diverse real-world benchmarks [2, 17, 36, 47, 49].

3 Method

Given multi-view images with calibrated cameras, our goal is to recover a precise SMPL-X body mesh (Fig. 2). Rather than directly regressing model parameters from pixels, we train a diffusion-based generator on large-scale synthetic data with pixel-perfect annotations (Sec. 3.1) to predict dense pixel-to-surface correspondences from each input view (Sec. 3.2). SMPL-X parameters are then

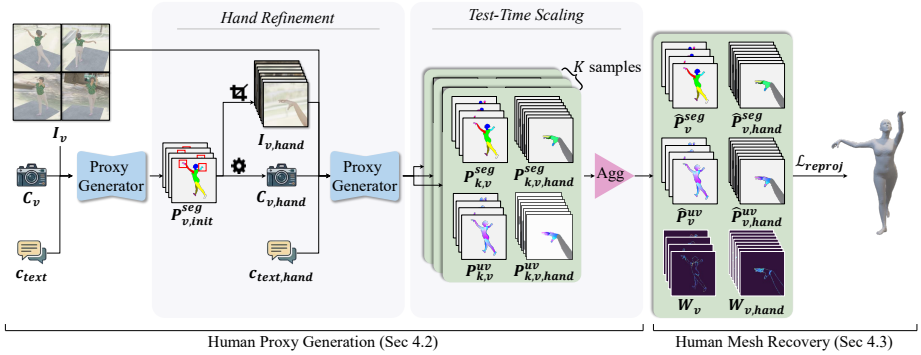

Fig. 2: Method overview. The proxy generator first produces initial body proxies; the segmentation component $\mathbf{P}^{\text{seg}}_{v,\text{init}}$ is then used to crop and enlarge hand regions for a second pass (hand refinement). Test-time scaling then draws K stochastic samples for both body and hand proxies, aggregates them, and derives per-pixel uncertainty weight maps $\mathbf{W}_v$. Finally, SMPL-X parameters are recovered by minimizing an uncertainty-weighted reprojection objective $\mathcal{L}_{\text{reproj}}$.

Fig. 3: Synthetic training data samples. Our dataset covers diverse poses, realistic occlusions, varied lighting, and clothing simulation, all with pixel-perfect SMPL-X proxy annotations (segmentation and UV).

recovered by minimizing a unified reprojection objective against these proxies (Sec. 3.3).

3.1 Synthetic Data Preparation

We train our model on a large-scale synthetic multi-view dataset with pixel-perfect SMPL-X annotations, eliminating annotation noise inherent in real-world datasets. A single model is trained on this dataset and used for all experiments without dataset-specific fine-tuning.

We construct our dataset by rendering clothed subjects from two sources: 67,650 from BEDLAM [3] driven by AMASS [34] motion sequences, and 37,837 from SynBody [53] driven by MPI-3DHP [36] and MoYo [47] poses, totaling

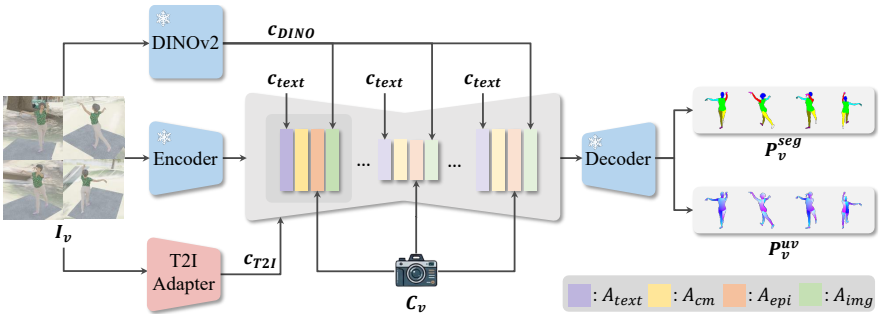

Fig. 4: Diffusion-based proxy generator architecture. Our model is built on Stable Diffusion 2.1 with a frozen UNet backbone, equipped with three conditioning signals ($\mathbf{c}_{\text{text}}$, $\mathbf{c}_{\text{T2I}}$, $\mathbf{c}_{\text{DINO}}$) and four trainable attention modules ($\mathcal{A}_{\text{text}}$, $\mathcal{A}_{\text{img}}$, $\mathcal{A}_{\text{cm}}$, $\mathcal{A}_{\text{epi}}$) for multi-view consistent proxy generation.

105,487 unique SMPL-X subjects. We only use training-set motions from MPI-3DHP and MoYo; their test-set poses and all motions from our evaluation datasets are excluded. All visual content (cameras, backgrounds, textures, lighting) is entirely synthetic, ensuring that evaluation is unaffected by our data generation process. Our rendering pipeline incorporates realistic occlusions from 7,953 object meshes in Amazon Berkeley Objects [8], diverse hairstyles from Hair20K [14], 863 HDR environment maps from Poly Haven [41] serving as both lighting and background, and physically-based clothing simulation [3] (Fig. 3). For each subject, we sample 8 cameras and render 1024×1024 RGB images with corresponding SMPL-X proxies, yielding 105,487 multi-view samples (843,896 images in total).

3.2 Human Proxy Generation

SMPL-X and proxy definition. SMPL-X [40] is a parametric 3D human mesh model with parameters $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{T}\}$: shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, pose $\boldsymbol{\theta}$, facial expression $\boldsymbol{\psi}$, and global translation $\mathbf{T} \in \mathbb{R}^3$. Each vertex carries a 2D uv coordinate $\mathbf{u} \in [0,1]^2$ on a predefined texture map partitioned by body parts.

We define SMPL-X proxy as a 2D dense representation establishing pixel-to-surface correspondences. For view v, the proxy $\mathbf{P}_v = (\mathbf{P}_v^{\text{seg}}, \mathbf{P}_v^{\text{uv}})$ consists of segmentation and UV components. Both are encoded as three-channel images to match the pre-trained VAE decoder of SD 2.1, which empirically yields better reconstruction quality than single-channel alternatives. $\mathbf{P}_v^{\text{seg}}$ assigns each semantic body part a unique RGB color. $\mathbf{P}_v^{\text{uv}}$ stores the uv coordinates in the first two channels with the third set to one. We render the textured mesh with perspective projection to obtain the proxy images.

Diffusion-based proxy generator. Our generator G_ϕ is built on Stable Diffusion 2.1 [45] with frozen UNet backbone (Fig. 4). Given multi-view images $\{I_v\}_{v=1}^N$ ($N \geq 1$) and perspective camera parameters $\{C_v\} = \{(\mathbf{K}_v, \mathbf{R}_v, \mathbf{t}_v)\}$

(no lens distortion), the model predicts SMPL-X proxies $\mathbf{P}_v = (\mathbf{P}_v^{\text{seg}}, \mathbf{P}_v^{\text{uv}}) \in \mathbb{R}^{256\times256\times3}$ encoding body part labels and surface coordinates.

Three conditioning signals drive the generation: text prompts $\mathbf{c}_{\text{text}}$ specify the target modality and body region (*e.g.*, "full body segmentation map" or "full body UV map"), determining which proxy is generated; T2I-Adapter [37] features $\mathbf{c}_{\text{T2I}} = \mathcal{E}_{\text{T2I}}(I_v)$ enforce pixel-level alignment; DINOv2 [38] tokens $\mathbf{c}_{\text{DINO}} = \mathcal{E}_{\text{DINO}}(I_v)$ provide pose and appearance priors. $\mathbf{c}_{\text{text}}$ and $\mathbf{c}_{\text{DINO}}$ are injected via text cross-attention $\mathcal{A}_{\text{text}}$ and image cross-attention $\mathcal{A}_{\text{img}}$, while $\mathbf{c}_{\text{T2I}}$ is added as residual features. For multi-modal and multi-view consistency, trainable cross-modality attention $\mathcal{A}_{\text{cm}}$ concatenates UV and segmentation tokens, and multi-view epipolar attention $\mathcal{A}_{\text{epi}}$ enforces geometric consistency via epipolar-constrained self-attention [23] with Plücker ray embeddings. Only the attention modules and T2I-Adapter are trained; the UNet and DINOv2 remain frozen.

The model is trained with the standard diffusion objective. Given ground-truth proxy $\mathbf{P}_v^*$, we encode it to latent $\mathbf{z}_0 = \mathcal{E}_{\text{VAE}}(\mathbf{P}_v^*)$ via the VAE encoder, sample timestep t and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and optimize:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},t,\mathbf{c}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right], \tag{1}$$

where $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$ and $\mathbf{c} = \{\mathbf{c}_{\text{text}}, \mathbf{c}_{\text{T2I}}, \mathbf{c}_{\text{DINO}}\}$. We train with a fixed budget of $N=4$ views per sample, where 2–4 views are full-body and the remaining slots are filled with hand-region crops (left/right randomly selected) from the same camera viewpoints. This enables joint body and hand training without architectural changes. At inference, we denoise a random latent $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and decode via VAE decoder $\mathcal{D}$ to obtain $\mathbf{P}_v$. The model generalizes to different view counts without fine-tuning.

Hand refinement. In full-body images I_v, hands occupy <1% of pixels, producing sparse and unreliable proxy predictions in these regions. We adopt a two-pass strategy: first inferring initial body proxies $\mathbf{P}_{v,\text{init}}$, then using the segmentation component $\mathbf{P}_{v,\text{init}}^{\text{seg}}$ to localize hand regions and create enlarged crops $I_{v,\text{hand}}$. When cropping, extrinsic parameters remain unchanged while intrinsics (focal length and principal point) are adjusted to produce hand-specific camera parameters $C_{v,\text{hand}}$. In the second pass, we treat hand crops as additional views with $C_{v,\text{hand}}$ and switch to a hand-specific text prompt $\mathbf{c}_{\text{text,hand}}$ that subdivides each hand into 12 parts (two palms and ten fingers), leveraging cross-view attention to produce refined hand proxies $\mathbf{P}_{v,\text{hand}}$. This coarse-to-fine strategy improves finger fidelity without modifying the network architecture.

Test-time scaling & uncertainty. Individual pixel predictions can still be erroneous in challenging regions such as self-occlusions, part boundaries, or visually ambiguous areas. We leverage diffusion stochasticity to identify and mitigate such errors: by drawing K stochastic samples $\{\mathbf{P}_{k,v}\}_{k=1}^K$ per view and measuring their disagreement, we obtain per-pixel uncertainty estimates that down-weight unreliable predictions during optimization.

For UV aggregation, we compute the pixel-wise median across samples to obtain a robust estimate:

$$\hat{\mathbf{P}}_v^{\text{uv}}(x) \;=\; \text{median}_{k=1..K}\big[\,\mathbf{P}_{k,v}^{\text{uv}}(x)\,\big], \tag{2}$$

and quantify uncertainty using the channel-wise sample variance averaged over three channels:

$$\mathbf{U}_v^{\mathrm{uv}}(x) \;=\; \frac{1}{3}\sum_{c=1}^{3} \mathrm{Var}_{k=1..K}\big[\,\mathbf{P}_{k,v}^{\mathrm{uv},(c)}(x)\,\big]. \tag{3}$$

For segmentation, we first quantize each sample $\mathbf{P}_{k,v}^{\mathrm{seg}}$ to the nearest color in a predefined palette $\mathcal{P}_{\mathrm{view}}$ (hand part palette for hand crops, body part palette otherwise). We then apply pixel-wise majority voting across K samples to obtain the aggregated segmentation $\hat{\mathbf{P}}_v^{\mathrm{seg}}$. The segmentation uncertainty measures how strongly the K samples agree: if the most-voted label at pixel x receives $n_{\max}(x)$ votes, the uncertainty is

$$\mathbf{U}_v^{\mathrm{seg}}(x) \;=\; \begin{cases} 1, & n_{\max}(x) \le \frac{K}{2}, \\ 2\left(1 - \dfrac{n_{\max}(x)}{K}\right), & \text{otherwise}, \end{cases} \tag{4}$$

which equals 1 (maximum) when no label achieves majority and decreases linearly to 0 as agreement reaches 100%.

The uncertainties modulate the fitting via a per-view weight map $\mathbf{W}_v \in \mathbb{R}^{256\times256}$, where each pixel x is assigned a reliability weight:

$$\mathbf{W}_v(x) \;=\; \big(1 - \mathbf{U}_v^{\mathrm{uv}}(x)\big)\,\big(1 - \mathbf{U}_v^{\mathrm{seg}}(x)\big). \tag{5}$$

This strategy provides a compute–accuracy trade-off through K without test-time adaptation of network weights.

3.3 Human Mesh Recovery

Our dense proxies provide far richer constraints than sparse keypoints: each foreground pixel carries a semantic part label and UV coordinate on the SMPL-X surface, offering dense coverage across the entire visible body.

Given proxies $\{\mathbf{P}_v\}$ from all views, we compute the reprojection loss over all foreground pixels. Let $\mathrm{fg}(v)$ denote foreground pixels in view v. For each pixel $x \in \mathrm{fg}(v)$, we map it to a 3D point on the SMPL-X surface via its proxy label and UV coordinate, and compute the L2 reprojection error $d(x)$ between the projected surface point and the original pixel (see Algorithm 1 in supplementary). The reprojection loss, weighted by the per-pixel reliability map $\mathbf{W}_v$ (Eq. (5)), is:

$$\mathcal{L}_{\mathrm{reproj}} = \sum_{v} \sum_{x\in\mathrm{fg}(v)} \mathbf{W}_v(x)\, d(x)^2. \tag{6}$$

When test-time scaling is disabled ($K=1$), all weights default to $\mathbf{W}_v(x)=1$.

Optimization. We optimize body and hand poses in axis-angle space, and shape $\boldsymbol{\beta}$ without explicit regularization. Body and hand poses are regularized by

DPoser-X [33], a diffusion-based whole-body pose prior, via a one-step denoising loss $\mathcal{L}_{\text{prior}}$. The total objective is:

$$\mathcal{L} = \mathcal{L}_{\text{reproj}} + \lambda_{\text{prior}} \mathcal{L}_{\text{prior}}, \quad \lambda_{\text{prior}} = 0.1. \tag{7}$$

We minimize $\mathcal{L}$ using Adam with stage-wise learning rate scheduling. See Sec. 4 for implementation details.

4 Experiments

4.1 Implementation Details

Proxy generator training. We trained with 4 views per sample using random full-body/hand crops and bbox augmentation. From SD-2.1 pre-trained weights, we optimized only the attention modules ($\mathcal{A}_{\text{text}}$, $\mathcal{A}_{\text{img}}$, $\mathcal{A}_{\text{cm}}$, $\mathcal{A}_{\text{epi}}$) and T2I-Adapter $\mathcal{E}_{\text{T2I}}$ while freezing the UNet backbone and DINOv2 $\mathcal{E}_{\text{DINO}}$. Training used batch size 2, Adam optimizer, learning rate 5×10^{-5}, for 30 epochs on 4× RTX 5090 GPUs (∼36 hours).
VAE decoder refinement. Stable Diffusion's pre-trained VAE decoder $\mathcal{D}$ may introduce quantization artifacts for proxy representations that require high numerical precision. We fine-tuned $\mathcal{D}$ with learning rate 1×10^{-6}, batch size 8, for 100K iterations (∼4 hours on 4× RTX 5090).
Inference. We generate proxies $\mathbf{P}_v$ for 12 views by default: 4 full-body views plus left/right hand crops for each. Inference involves two passes on a single RTX 5090: first generating 4 full-body proxies to localize hand regions (∼3s), then generating all 12 proxies including hand crops (∼15s). With test-time scaling, proxy generation takes $K \times 15$s (default $K = 3$). Mesh fitting adds ∼50s without hand refinement or ∼60s with it.
Mesh fitting. We optimize $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\mathbf{T}$ from Θ, along with a global scale parameter, using Adam in a coarse-to-fine schedule that first solves global placement, then body pose and shape, and finally hand articulations. We advance to the next stage when the loss converges, *i.e.*, the relative decrease between consecutive iterations falls below a per-stage threshold (1%–10%). Facial expression $\boldsymbol{\psi}$ is not optimized as our focus is on body and hand reconstruction.

4.2 Datasets and Baselines

Datasets. We evaluate on five real-world datasets: 3DHP [36], BEHAVE [2], RICH [17], MoYo [47], and 4D-DRESS [49], covering studio capture, human-object interaction, outdoor scenes, challenging poses, and loose clothing. We additionally test on 4D-DRESS with random crops (4D-DRESS partial) to evaluate robustness to partial observations. All real datasets are unseen during training; a single model is used without dataset-specific fine-tuning.
Baselines. We compare against six baselines spanning three categories: *single-view end-to-end*—SMPLest-X [56] (trained on large-scale diverse data) and Human3R [5] (extending CUT3R for joint human-scene recovery); *multi-view end-*

Table 1: Quantitative comparison on five real-world datasets. * indicates the method was trained on that specific dataset. For single-view methods, the best single-camera result is reported. Three configurations: "w/o HR, TTS" (4 full-body views only), "w/o TTS" (+ hand refinement), "Ours" (full model). Best / second best.

	3dhp				*rich*				*behave*			
Method	PA-MPJPE	MPJPE	PA-MPVPE	MPVPE	PA-MPJPE	MPJPE	PA-MPVPE	MPVPE	PA-MPJPE	MPJPE	PA-MPVPE	MPVPE
SMPLest-X [56]	32.5*	48.1*	46.6*	62.3*	31.8*	67.6*	40.3*	81.1*	30.0*	48.6*	43.5*	63.6*
Human3R [5]	53.1	84.3	67.7	104.1	52.1	101.1	64.4	116.9	38.3	85.0	53.8	102.3
U-HMR [28]	67.3*	142.9*	78.5*	164.2*	55.5	134.9	67.8	159.3	46.1	118.7	51.9	135.0
MUC [59]	40.2	-	51.3	-	35.7*	-	42.7*	-	28.7	-	41.8	-
HeatFormer [35]	29.2*	49.9*	34.7*	53.3*	40.1	80.1	52.9	93.8	33.4	70.2	47.7	80.5
EasyMoCap [1]	34.3	43.7	43.2	51.5	31.2	49.8	44.2	60.3	21.8	29.6	35.1	40.4
Ours (w/o HR, TTS)	33.4	42.0	43.4	50.5	25.6	36.6	37.5	46.5	23.6	33.7	33.7	43.1
Ours (w/o TTS)	33.1	41.4	45.3	51.5	24.5	34.6	29.3	36.2	23.5	33.5	32.2	40.7
Ours	32.8	41.1	44.7	50.9	24.2	33.7	28.9	35.1	22.8	32.7	31.3	39.9

	moyo				*4ddress*				*4ddress-partial*			
Method	PA-MPJPE	MPJPE	PA-MPVPE	MPVPE	PA-MPJPE	MPJPE	PA-MPVPE	MPVPE	PA-MPJPE	MPJPE	PA-MPVPE	MPVPE
SMPLest-X [56]	45.6*	63.3*	62.1*	81.4*	34.8	52.1	52.4	70.2	72.2	101.7	111.3	140.6
Human3R [5]	65.0	87.8	81.6	107.6	30.0	52.4	42.4	66.7	72.3	117.6	103.5	153.9
U-HMR [28]	88.3	177.7	104.9	209.7	42.1	80.3	52.6	98.8	65.2	136.1	87.2	175.2
MUC [59]	62.8	-	76.0	-	33.2	-	49.2	-	60.1	-	92.9	-
HeatFormer [35]	67.3	126.9	79.6	123.6	43.4	73.3	61.6	90.2	143.8	291.6	175.2	327.0
EasyMoCap [1]	31.4	41.2	44.8	51.6	19.9	24.9	31.8	35.3	48.2	94.8	69.8	119.4
Ours (w/o HR, TTS)	29.4	33.6	41.8	44.9	17.6	22.4	30.4	33.4	29.6	34.4	49.6	52.0
Ours (w/o TTS)	29.2	33.0	36.3	39.6	17.0	20.8	24.0	26.2	28.1	32.6	42.3	44.4
Ours	28.3	32.4	35.2	38.8	16.6	20.5	23.3	25.6	27.6	31.6	41.3	43.1

to-end—U-HMR [28] (decoupled camera and body estimation), MUC [59] (prediction fusion), and HeatFormer [35] (neural optimization with heatmaps); and *fitting-based*—EasyMoCap [1] (sparse keypoint fitting).

4.3 Quantitative Results

Evaluation protocol. We evaluate on official test/validation splits with a fixed set of 4 views per dataset. On 3DHP [36], we follow HeatFormer [35] and use subject S8 with views 0/2/7/8, filtering invalid masks as in their protocol. BEHAVE [2] uses the official test split with all 4 views; RICH [17] uses the first 4 valid views from the test split; MoYo [47] uses the validation split with views 1/3/4/5. Since 4D-DRESS [49] lacks an official split, we evaluate on the full dataset with all 4 views. To reduce temporal redundancy, we sample every 5th frame [3, 26, 42]; for 4D-DRESS we sample every 50th frame to keep evaluation tractable. We report MPJPE, MPVPE, and their Procrustes-aligned variants in millimeters. For single-view baselines, we evaluate each camera view independently after root alignment and report the single view with the lowest MPVPE per dataset, giving these methods a favorable comparison.

As shown in Tab. 1, our method achieves state-of-the-art performance across all six evaluation settings. We organize our analysis around three observations:

Fitting-based vs. end-to-end methods. End-to-end methods generally exhibit higher errors than fitting-based approaches on absolute positioning metrics. For example, on RICH the best end-to-end MPVPE is 81.1 mm (SMPLest-X), whereas EasyMoCap achieves 60.3 mm and our full model 35.1 mm, confirming that constructing accurate intermediate representations and fitting against them is more effective than direct regression. Notably, even our base model (w/o HR, TTS)—using the same 4-view input as multi-view baselines—already surpasses all end-to-end methods on most metrics.

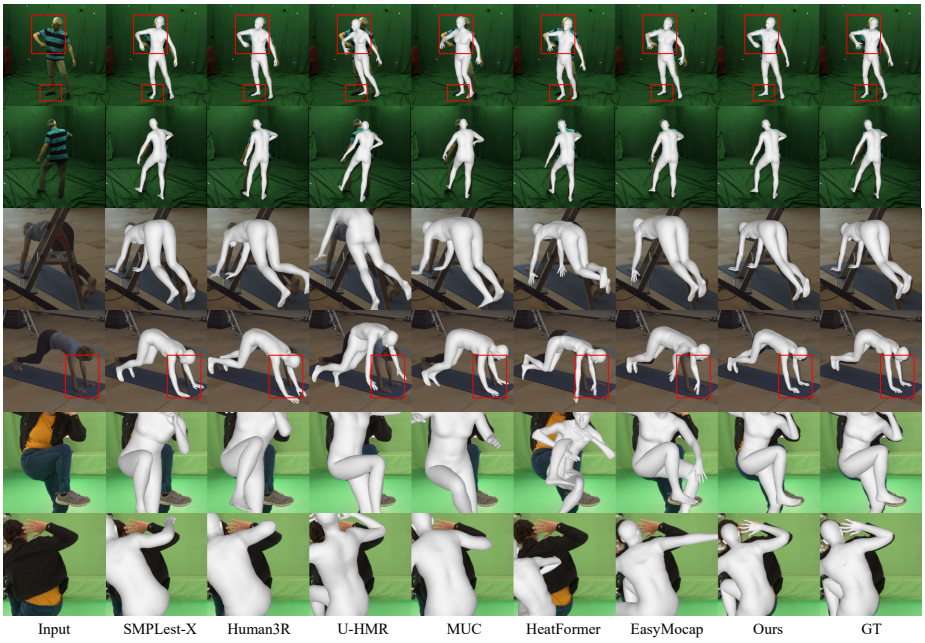

Fig. 5: Qualitative comparison across three scenarios: 3DHP (rows 1–2), RICH (rows 3–4), and 4D-DRESS partial (rows 5–6). Red boxes highlight notable misalignment. Our method maintains tight image-mesh correspondence across all settings.

Cross-dataset generalization. Existing multi-view methods are limited by scarce real multi-view data and tend to excel only on their training domains—*e.g.*, HeatFormer [35] achieves competitive MPVPE (53.3) on 3DHP where it is trained, but drops to 123.6 on MoYo and 327.0 on 4D-DRESS partial, falling behind even single-view methods. Our large-scale synthetic dataset (105K multi-view samples) combines data diversity with multi-view geometric supervision, enabling strong performance across all benchmarks without real training data.

Dense proxies vs. sparse keypoints. EasyMoCap [1] is also fitting-based but relies on sparse keypoints. The advantage of dense proxies manifests in two aspects. (1) *Precision*: our proxy maps each foreground pixel to a specific surface point, constraining the full mesh surface rather than just joint locations. We outperform EasyMoCap in MPVPE on all six settings (*e.g.*, 35.1 vs. 60.3 on RICH, 25.6 vs. 35.3 on 4D-DRESS). (2) *Robustness*: EasyMoCap performs well where keypoint detection is reliable (3DHP, BEHAVE) but degrades under complex lighting (RICH), extreme poses (MoYo), and loose clothing (4D-DRESS). The gap widens with partial views: EasyMoCap's MPVPE increases by 238% from 4D-DRESS to 4D-DRESS partial (35.3→119.4), while ours increases by only 68% (25.6→43.1), as dense constraints across the visible surface make the fitting resilient to partial occlusion.

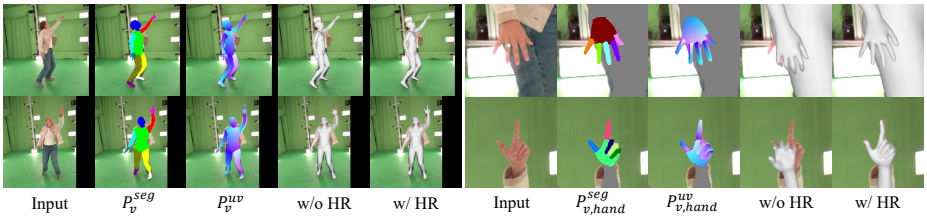

Fig. 6: Qualitative comparison of hand refinement. Hand refinement improves fitting quality and produces accurate finger details.

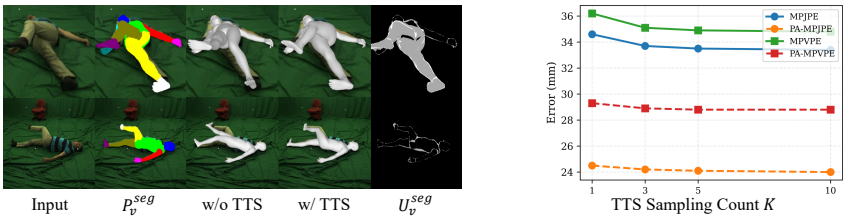

Fig. 7: Test-time scaling with uncertainty weighting improves robustness by down-weighting unreliable predictions and recovering correct poses from erroneous proxy outputs (left). Increasing K consistently improves reconstruction quality (right).

4.4 Qualitative Results

Figure 5 presents qualitative comparisons across three representative scenarios. On 3DHP (rows 1–2), the ground-truth meshes show visible offset from the images in certain regions (red boxes in row 1); methods trained on this dataset (HeatFormer, SMPLest-X) tend to reproduce similar alignment patterns, suggesting that their metrics partly reflect annotation biases. Our method achieves tighter visual alignment despite lower scores on this benchmark. On RICH (rows 3–4), the cluttered real-world environment misleads EasyMoCap's sparse keypoint detection, resulting in degraded fitting quality, while our dense proxies remain robust to such visual distractors. On 4D-DRESS partial (rows 5–6), most methods degrade severely with limited visibility, whereas our dense surface constraints remain effective even under heavy occlusion.

4.5 Ablation Studies

Hand refinement. Comparing "w/o HR, TTS" and "w/o TTS" in Tab. 1 reveals that hand refinement consistently improves accuracy. In full-body images, hands occupy <1% of pixels; hand refinement produces clearer proxies that better localize hand position, anchoring arm orientation and improving full-body accuracy. Figure 6 shows that refinement produces hand proxies with more accurate finger poses and less part ambiguity.

Test-time scaling and uncertainty weighting. Figure 7 illustrates: a single prediction may swap the left and right leg labels, but across K stochastic

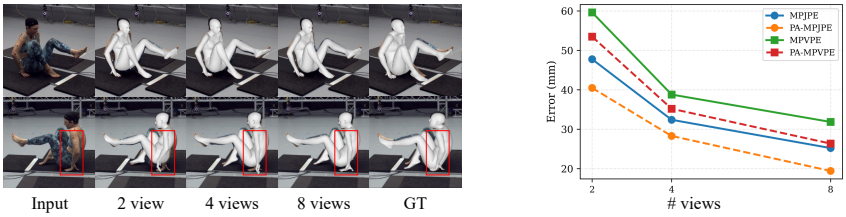

Fig. 8: Our method benefits from increasing view counts, with performance improving progressively from 2-view to 8-view configurations.

Table 2: Ablation study on BEHAVE. Top: removing individual network modules. Bottom: inference without ground-truth cameras.

Configuration	PA-MPJPE	MPJPE	PA-MPVPE	MPVPE
w/o DINOv2	37.3	44.3	53.7	60.2
w/o T2I-Adapter	28.5	41.5	36.7	48.4
w/o $\mathcal{A}_{\text{text}}$	26.5	35.2	35.3	43.3
w/o $\mathcal{A}_{\text{epi}}$	25.7	33.7	35.0	41.9
w/o $\mathcal{A}_{\text{cm}}$	26.3	33.8	35.4	41.9
w/o GT camera	25.9	–	37.3	–
Ours	**22.8**	**32.7**	**31.3**	**39.9**

samples some predictions are correct while others are not; this disagreement is captured by $\mathbf{U}_v^{\text{seg}}$, which assigns high uncertainty to inconsistent regions. Fitting then down-weights these pixels and relies on more consistent views to recover the correct configuration. Increasing K generally improves performance (Fig. 7, right); we use $K=3$ as default.

Number of input views. Our method supports flexible view counts without retraining. As shown in Fig. 8 (evaluated on MoYo), even two views already produce reasonable reconstructions. Adding more views further improves the recovery of occluded details—*e.g.*, arms behind the torso and hands near the hip (red boxes) become progressively more accurate as additional viewpoints increase their visibility.

Network module contributions. Table 2 ablates individual network modules on BEHAVE. We independently remove DINOv2, T2I-Adapter, and attention modules ($\mathcal{A}_{\text{text}}$, $\mathcal{A}_{\text{epi}}$, $\mathcal{A}_{\text{cm}}$). Each module contributes to overall performance, with the full model achieving the best results.

Inference without camera calibration. While our main results assume calibrated cameras, real-world scenarios often lack ground-truth camera parameters. We test a camera-free variant on BEHAVE: we predict camera parameters using VGGT [48], then generate proxies with these predicted cameras. During mesh fitting, we jointly optimize camera parameters alongside body pose and shape to compensate for prediction inaccuracy. As shown in Tab. 2, our method achieves competitive performance with only moderate degradation.

Why diffusion-based proxy generation? A natural alternative is a discriminative model such as DensePose [13]. We fine-tune DensePose on our synthetic

Table 3: Comparison with DensePose [13] and its fine-tuned variant on MoYo.

Method	PA-MPJPE	MPJPE	PA-MPVPE	MPVPE
DensePose	67.6	101.1	90.4	127.8
DP-FT (60ep)	66.5	88.8	84.5	111.4
Ours	**28.3**	**32.4**	**35.2**	**38.8**

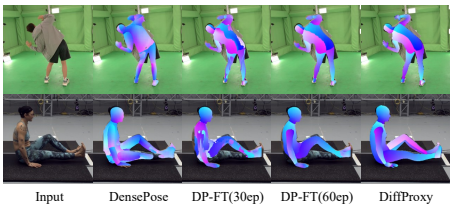

Fig. 9: Comparison with DensePose [13] and its fine-tuned variants (DP-FT). Fine-tuning improves standard poses (left) but provides limited benefit on challenging ones (right). DiffProxy produces accurate proxies in both cases.

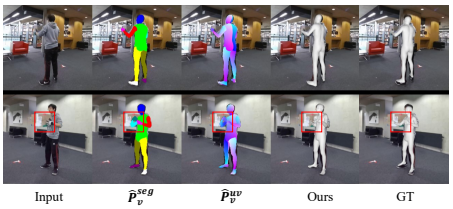

Fig. 10: A typical failure case on BEHAVE. The proxy predicts correct dense correspondences, but mesh fitting converges to a local optimum with inverted joint orientations—the pose appears correct in 2D but is flipped in depth.

dataset for 30 and 60 epochs and run our fitting pipeline on its predictions. Figure 9 shows qualitative results; Tab. 3 reports the better variant (60 epochs). Fine-tuning improves DensePose on standard poses, confirming that our synthetic data provides effective supervision, but even with extended training DensePose still produces fragmented predictions on challenging poses. Our diffusion-based generator leverages the visual priors of Stable Diffusion 2.1 for robust generalization across diverse image compositions, naturally supporting hand refinement on enlarged crops, and enabling test-time scaling for uncertainty estimation via stochastic sampling.

5 Limitation and Future Works

While DiffProxy achieves state-of-the-art performance, several limitations remain. **Fitting local optima**: The 2D reprojection objective can converge to local optima where joint configurations appear correct in the image plane but are inverted in 3D, as illustrated in Fig. 10. Incorporating complementary 3D constraints (*e.g.*, triangulating correspondences across views) could help resolve such depth ambiguities. **Inference speed**: Without hand refinement or test-time scaling, our method takes ∼53s per subject (3s proxy generation + 50s fitting), comparable to EasyMoCap (∼40s) while already achieving superior precision and robustness (Tab. 1, "w/o HR, TTS"). Hand refinement and test-time scaling provide further gains at additional cost. Future work could accelerate fitting by initializing with end-to-end predictions, and speed up proxy generation via consistency models [44] or diffusion distillation. **Multi-view requirement**: Our method requires multiple views for reliable results, as single-view performance

suffers from depth ambiguity. **Single-subject**: Extension to multi-person scenarios is straightforward by incorporating per-instance segmentation, with the primary challenge being cross-view identity association.

6 Conclusion

We showed that the core of fitting-based human mesh recovery—predicting dense pixel-to-surface correspondences from images—is an image-to-image translation task well-suited for modern diffusion models. By training on large-scale synthetic data with pixel-perfect annotations and leveraging the visual priors of Stable Diffusion, DiffProxy overcomes the two bottlenecks that hindered prior dense-proxy approaches: noisy manual labels and limited generalization. Combined with hand refinement for fine-grained detail and test-time scaling for uncertainty-guided fitting, the resulting dense proxies provide uniform surface constraints that enable precise fitting. Trained exclusively on synthetic data without dataset-specific tuning, DiffProxy achieves state-of-the-art performance across five diverse real-world benchmarks, suggesting that the dense proxy route is a highly effective paradigm for precision-oriented tasks.

7 Acknowledgments

This work was supported by the National Science Fund of China under Grant Nos. U24A20330 and 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. Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (jun 2022)
3. Black, M.J., Patel, P., Tesch, J., Yang, J.: Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8726–8737 (2023)
4. 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: European Conference on Computer Vision. pp. 561–578. Springer (2016)
5. Chen, Y., Chen, X., Xue, Y., Chen, A., Xiu, Y., Pons-Moll, G.: Human3r: Everyone everywhere all at once. arXiv preprint arXiv:2510.06219 (2025)
6. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., et al.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19982–19993 (2023)

7. Cho, H., Kim, J.: Generative approach for probabilistic human mesh recovery using diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4183–4188 (2023)
8. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., et al.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21126–21136 (2022)
9. Foo, L.G., Gong, J., Rahmani, H., Liu, J.: Distribution-aligned diffusion for human mesh recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9221–9232 (2023)
10. Gao, J., Zheng, C., Jeni, L.A., Erickson, Z.: Disrt-in-bed: Diffusion-based sim-to-real transfer framework for in-bed human mesh recovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1829–1838 (2025)
11. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14783–14794 (2023)
12. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10884–10894 (2019)
13. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7297–7306 (2018)
14. 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 (2025)
15. Heo, J., Wang, K.C., Liu, K., Yeung-Levy, S.: Motion diffusion-guided 3d global hmr from a dynamic camera. arXiv preprint arXiv:2411.10582 (2024)
16. Hewitt, C., Saleh, F., Aliakbarian, S., Petikam, L., Rezaeifar, S., Florentin, L., Hosenie, Z., Cashman, T.J., Valentin, J., Cosker, D., Baltrušaitis, T.: Look ma, no markers: Holistic performance capture without the hassle. ACM Trans. Graph. (2024)
17. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13274–13285 (Jun 2022)
18. 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. arXiv preprint arXiv:2412.03632 (2024)
19. Huang, Z., Wen, H., Dong, J., Wang, Y., Li, Y., Chen, X., Cao, Y.P., Liang, D., Qiao, Y., Dai, B., et al.: Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9784–9794 (2024)
20. Jeong, Y., Lee, J., Kim, C., Cho, M., Lee, D.: Nvs-adapter: Plug-and-play novel view synthesis from a single image. In: European Conference on Computer Vision. pp. 449–466. Springer (2024)
21. Jia, K., Zhang, H., An, L., Liu, Y.: Delving deep into pixel alignment feature for accurate multi-view human mesh recovery. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 989–997 (2023)

22. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5614–5623 (2019)
23. Kant, Y., Siarohin, A., Wu, Z., Vasilkovsky, M., Qian, G., Ren, J., Guler, R.A., Ghanem, B., Tulyakov, S., Gilitschenski, I.: Spad: Spatially aware multi-view diffusers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10026–10038 (2024)
24. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9492–9502 (2024)
25. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2252–2261 (2019)
26. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: CVPR (2017)
27. Lê, E.T., Kakolyris, A., Koutras, P., Tam, H., Skordos, E., Papandreou, G., Güler, R.A., Kokkinos, I.: Meshpose: Unifying densepose and 3d body mesh reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2405–2414 (2024)
28. Li, X., Meng, M., Wu, Z., Chen, T., Yang, F., Shen, D.: Human mesh recovery from arbitrary multi-view images. arXiv preprint arXiv:2403.12434 (2024)
29. 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. In: Advances in Neural Information Processing Systems. vol. 36, pp. 25268–25280 (2023)
30. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1954–1963 (2021)
31. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. In: ICLR (2024)
32. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
33. Lu, J., Lin, J., Dou, H., Zeng, A., Deng, Y., Liu, X., Cai, Z., Yang, L., Zhang, Y., Wang, H., Liu, Z.: Dposer-x: Diffusion model as robust 3d whole-body human pose prior. arXiv preprint arXiv:2508.00599 (2025)
34. 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. pp. 5442–5451 (Oct 2019)
35. Matsubara, Y., Nishino, K.: Heatformer: A neural optimizer for multiview human mesh recovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6415–6424 (2025)
36. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3D Vision (3DV), 2017 Fifth International Conference on. IEEE (2017). https://doi.org/10.1109/3dv.2017.00064, http://gvv.mpi-inf.mpg.de/3dhp_dataset
37. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)

38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
39. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13468–13478 (2021)
40. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10975–10985 (2019)
41. Poly Haven: Poly haven: The public 3d asset library. `https://polyhaven.com/` (2024), accessed 2025-11-12
42. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2d and 3d pose detection in natural images. IEEE TPAMI (2019)
43. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)
44. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: International Conference on Machine Learning (2023)
45. Stability AI: Stable diffusion v2.1 release. `https://stability.ai/news/stablediffusion2-1-release7-dec-2022` (2022), accessed 2025-10-21
46. Stathopoulos, A., Han, L., Metaxas, D.: Score-guided diffusion for 3d human recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 906–915 (2024)
47. Tripathi, S., Müller, L., Huang, C.H.P., Omid, T., Black, M.J., Tzionas, D.: 3D human pose estimation via intuitive physics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4713–4725 (2023), `https://ipman.is.tue.mpg.de`
48. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
49. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
50. Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C.: What matters when repurposing diffusion models for general dense perception tasks? arXiv preprint arXiv:2403.06090 (2024)
51. Xu, Y., Zhu, S.C., Tung, T.: Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7760–7770 (2019)
52. Yang, J., Cheng, Z., Duan, Y., Ji, P., Li, H.: Consistnet: Enforcing 3d consistency for multi-view images diffusion. arXiv (2023)
53. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20282–20292 (2023)

54. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 469–480 (2023)
55. Yin, W., Cai, Z., Wang, R., Wang, F., Wei, C., Mei, H., Xiao, W., Yang, Z., Sun, Q., Yamashita, A., Yang, L., Liu, Z.: Whac: World-grounded humans and cameras. In: European Conference on Computer Vision. pp. 20–37. Springer (2024)
56. Yin, W., Cai, Z., Wang, R., Zeng, A., Wei, C., Sun, Q., Mei, H., Wang, Y., Pang, H.E., Zhang, M., et al.: Smplest-x: Ultimate scaling for expressive human pose and shape estimation. arXiv preprint arXiv:2501.09782 (2025)
57. Zeng, W., Ouyang, W., Luo, P., Liu, W., Wang, X.: 3d human mesh regression with dense correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7054–7063 (2020)
58. Zheng, C., Liu, X., Peng, Q., Wu, T., Wang, P., Chen, C.: Diffmesh: A motion-aware diffusion framework for human mesh recovery from videos. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4891–4901. IEEE (2025)
59. Zhu, Y., Wang, S., Xu, M., Zhuang, Z., Wang, Z., Wang, K., Zhang, H., Wang, Q.: Muc: Mixture of uncalibrated cameras for robust 3d human body reconstruction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 11040–11048 (2025)