This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SAM 3D Body: Robust Full-Body Human Mesh Recovery

Xitong Yang[1,*], Devansh Kukreja[1,*], Don Pinkus[1,*], Anushka Sagar[1],
Taosha Fan[1], Jinhyung Park[1,2†], Soyong Shin[1,2†], Jinkun Cao[1], Jiawei Liu[1],
Nicolas Ugrinovic[1], Matt Feiszli[1,§], Jitendra Malik[1,§], Piotr Dollar[1,§], Kris Kitani[1,§]

[1]Meta Superintelligence Labs, [2] Carnegie Mellon University



Figure 1. Full-body human mesh recovery results using SAM 3D Body (3DB). Our model demonstrates robust performance in estimating challenging poses across diverse viewpoints and produces accurate body and hand pose estimations within a unified framework.

Abstract

We introduce SAM 3D Body (3DB), a promptable model for single-image full-body 3D human mesh recovery (HMR) that demonstrates state-of-the-art performance, with strong generalization and consistent accuracy in diverse in-the-wild conditions. 3DB estimates the human pose of the body, feet, and hands. It is the first model to use a new parametric mesh representation, Momentum Human Rig (MHR), which decouples skeletal structure and surface shape. 3DB employs an encoder–decoder architecture and supports auxiliary prompts, including 2D keypoints and masks, enabling user-guided inference similar to the SAM family of models. We derive high-quality annotations from a multi-stage annotation pipeline that uses various combinations of manual keypoint annotation, differentiable optimization, multi-view geometry, and dense keypoint detection. Our data engine efficiently selects and processes data to ensure data diversity, collecting unusual poses and rare imaging conditions. We present a new evaluation dataset organized by pose and appearance categories, enabling nuanced analysis of model behavior. Our experiments demonstrate superior generalization and substantial improvements over prior methods in both qualitative user preference studies and traditional quantitative analysis. Both 3DB and MHR are open-source.

*Core contributor § Project lead
†Work done during the internship at Meta

1. Introduction

Estimating 3D human pose (skeleton structure) and shape (soft body tissue) from images is an essential capability for vision and embodied AI systems to understand and interact with people. Despite notable progress in human mesh recovery (HMR) [7, 9, 33, 51, 53], existing approaches still exhibit unsatisfactory robustness when applied to in-the-wild images, which limits their applicability to real-world scenarios such as robotics [32, 37, 48] and biomechanics [36]. In particular, current models often fail on individuals presenting challenging poses, severe occlusion, or captured from uncommon viewpoints. They also struggle to reliably estimate both the overall body pose and the fine details of the hands / feet in a unified full-body framework.

We argue that the primary challenges in developing a robust full-body human mesh recovery model stem from both the data and model aspects. First, collecting large-scale and diverse human pose datasets with high-quality mesh annotations is inherently difficult and computationally costly. Most existing datasets either suffer from low pose diversity due to laboratory capture settings [4, 12, 13] or from low mesh quality resulting from pseudo-labeling [1, 49]. Second, current HMR architectures do not address the distinct optimization mechanisms required for body and hand pose estimation, nor do they incorporate effective training strategies to handle ambiguity from monocular images.

In this work, we present SAM 3D Body (3DB), a robust

full-body HMR model fueled by large-scale, high-quality human pose data curated by our data engine.

Robust Full-body HMR Model. We make three main contributions to improve model performance on both body and hand pose estimation. (i) We propose a novel promptable encoder–decoder architecture [17, 39] that enables the model to condition on optional 2D keypoints, masks or camera information for controllable pose estimation. This promptable design naturally facilitates interactive guidance in ambiguous or challenging scenarios during training, and provides a coherent approach to integrate hand and body predictions. (ii) Our model utilizes a shared image encoder and two separate decoders for the body and hands. This two-way-decoder design effectively alleviates conflicts in optimizing body and hand pose estimation, which arise from differences in input resolution, camera estimation, and supervision objectives. (iii) Unlike most prior work that relies on the SMPL [26] human mesh model, we build 3DB on a new parametric mesh representation, MHR [8], which decouples skeletal pose and body shape, providing richer control and interpretability for full-body reconstruction.

Data Engine for Diverse Human Pose and High-quality Annotation. HMR methods have increasingly turned to large-scale training data for higher performance [3, 9, 55]. However, high-quality 3D supervision remains scarce, and existing in-the-wild datasets are still limited in scale and diversity. To this end, we design a new data creation pipeline that features: (i) *Data Quality*: Our annotation pipeline combines various combinations of components such as geometric constraints, parametric priors, and dense keypoint regression, which automatically yields high-quality 3D human mesh annotations. (ii) *Data Quantity*: We curate data from large licensed stock photo repositories, multiple multi-view capture datasets, and synthetic data. We create a large scale of **7 million** images with high-quality annotation. (iii) *Data Diversity*: Our data is diversified using a VLM-based data engine that mines for in-the-wild challenging images and routes them for annotation. This ensures coverage of rare poses, difficult viewpoints, and varied appearances, providing a more diverse dataset for supervision.

Together, the data engine and full-body HMR model enable 3DB to recover high-fidelity full-body human meshes from a single image. 3DB achieves state-of-the-art performance across both body and hand pose estimation. Extensive experiments demonstrate that 3DB consistently outperforms prior HMR methods on standard metrics, generalizes better to unseen datasets, and is preferred by users in a study of $7,800$ participants with a significant $5:1$ win rate. To our knowledge, it is the first single model that delivers the **best performance to body-specialized models and comparable performance to hand-specialized models**, while providing interactive control and strong robustness under challenging poses and in-the-wild scenarios.

2. Related Work

Human Mesh Models: The most widely used human mesh model is SMPL [26], which parameterizes human body into pose and shape. SMPL-X [34] goes further to include hands (MANO [40]) and faces (FLAME [21]). SMPL models intertwine the skeletal structure and soft-tissue mass within the *shape space*, which can limit interpretability and controllability. Alternatively, Momentum Human Rig [8], an enhancement of ATLAS [31], explicitly decouples the skeletal structure and body shape, and we adopt it as our representation of the human body.

Human Mesh Recovery (HMR): Early HMR methods like HMR 2.0 [9] were *body-only* methods that predicted the body without articulated hands or feet [7, 18, 22]. Instead, 3DB follows the more recent paradigm of full-body methods [2, 3, 5, 41, 53] that estimate *body+hands+feet*. There are also part-specific hand mesh recovery methods [35, 38] that only estimate the pose and shape of the hands, which usually have more accurate performance compared to full-body methods. Generative methods like UniHand [47] also generate compelling results with a unified diffusion model. In contrast, 3DB shows strong performance on both hand and full-body estimation with a feedfoward framework.

Promptable Inference: Promptable inference, popularized by the SAM family [17, 39], enables user or system-provided prompts (such as 2D keypoints or masks) to guide model predictions. Similarly to [53], our approach supports various prompt types, including 2D keypoints and masks, and by integrating prompt tokens directly into the transformer architecture, enables user-guided mesh recovery.

Data Quality and Annotation Pipelines: A major bottleneck in HMR is the quality of training data. Many datasets rely on pseudo-ground-truth (pGT) meshes obtained from monocular fitting [14, 18], which often contain systematic errors in pose, shape, and camera parameters [33]. Recent work [7, 51] highlights the impact of annotation noise on reported metrics and generalization. To address this, multi-view datasets [16, 28, 29] and synthetic data have been used in our work to provide higher-fidelity supervision. In all, our method builds on these insights by employing a scalable data engine that mines challenging cases and a multi-stage annotation pipeline that combines dense keypoint detection and robust optimization.

3. 3DB Model Architecture

Our goal is to recover 3D human meshes (*i.e.*, MHR parameters) accurately, robustly and interactively from a single image. To this end, we design 3DB as a promptable encoder–decoder architecture (see Figure 2) with a rich set of prompt tokens. 3DB is designed to be *interactive* as it can accept 2D keypoints or masks, allowing users or downstream systems to guide inference.

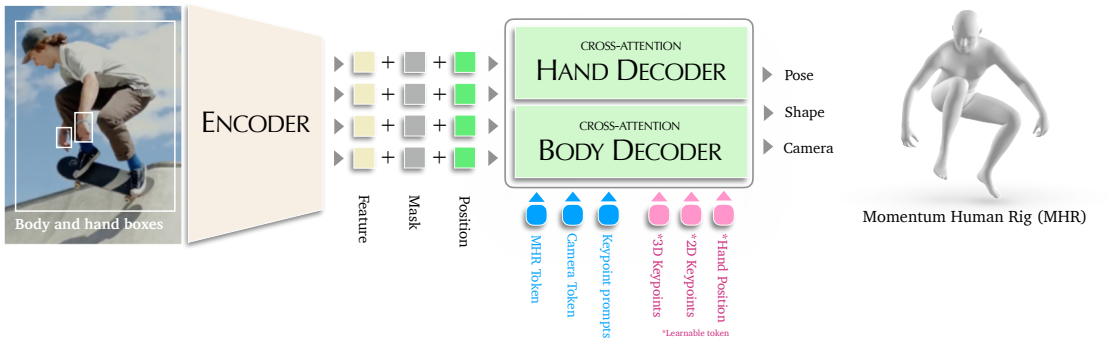


Figure 2. SAM 3D Body Model Architecture. We employ a promptable encoder–decoder architecture with a shared image encoder and separate decoders for body and hand pose estimation.

3.1. Image Encoder

The human-cropped image I is normalized and passed through a vision backbone to produce a dense feature map F. An optional set of hand crops I_{hand} can also be provided to obtain hand crop feature maps F_{hand}:

$$F = \text{ImgEncoder}(I), \tag{1}$$

$$F_{\text{hand}} = \text{ImgEncoder}(I_{\text{hand}}). \tag{2}$$

3DB considers two optional prompts: 2D keypoints and segmentation masks. Keypoint prompts are encoded by positional encodings summed with learned embeddings and are provided as additional tokens for the pose decoder. Mask prompts are embedded using convolutions and summed element-wise with the image embedding [17].

3.2. Decoder Tokens

3DB has two decoders: The body decoder outputs the full-body human rig and an optional hand decoder can provide enhanced hand pose results. The pose decoders take a set of *query tokens* as input to predict the parameters of MHR and camera parameters. There are four types of query tokens: MHR+camera, 2D keypoint prompt, auxiliary 2D/3D keypoint tokens and optional hand position tokens.

MHR+Camera Token: The initial estimate of MHR and (optionally) camera parameters is embedded as a learnable token for MHR parameter estimation:

$$T_{\text{pose}} = \text{RigEncoder}(E_{\text{init}}) \in \mathbb{R}^{1\times D}, \tag{3}$$

$$E_{\text{init}} \in \mathbb{R}^{d_{\text{init}}}. \tag{4}$$

2D Keypoint Prompt Tokens: If 2D keypoint prompts K are provided (*e.g.*, from a user or detector), they are encoded as:

$$T_{\text{prompt}} = \text{PromptEncoder}(K) \in \mathbb{R}^{N\times D}, \tag{5}$$

$$K \in \mathbb{R}^{N\times 3}, \tag{6}$$

where each keypoint is represented by (x, y, label).

Hand Position Tokens: The hand token, $T_{\text{hand}} \in \mathbb{R}^{2\times D}$, is used in the body decoder to locate the hand positions inside the human images. This set of tokens is optional, without which 3DB can still produce a full-body human rig because the output from body decoder already includes hands.

Auxiliary Keypoint Tokens: To further enhance interactivity and model capacity, we include learnable tokens for all 2D and 3D keypoints.

$$T_{\text{keypoint2D}} \in \mathbb{R}^{J_{2D}\times D}, \tag{7}$$

$$T_{\text{keypoint3D}} \in \mathbb{R}^{J_{3D}\times D}. \tag{8}$$

These tokens allow the model to reason about specific joints and support downstream tasks such as keypoint prediction or uncertainty estimation.

3.3. MHR Decoder

All tokens are concatenated to form the full set of queries:

$$T = [T_{\text{pose}},\ T_{\text{prompt}},\ T_{\text{keypoint2D}},\ T_{\text{keypoint3D}},\ T_{\text{hand}}] \tag{9}$$

This flexible assembly enables the model to operate in both fully automatic and user-guided modes, adapting to the available prompts. The body decoder attends to both the query tokens T, the full-body image features F,

$$O = \text{Decoder}(T, F) \in \mathbb{R}^{(3+N+J_{2D}+J_{3D})\times D}. \tag{10}$$

Through cross-attention, the body decoder fuses prompt information with visual context, enabling robust and editable mesh recovery. Optionally, the hand decoder can take the same prompt information while attends to the hand crop features F_{hand} to provide another output token O_{hand}.

The first output token of O is passed through an MLP to regress the final mesh parameters: $\theta = \text{MLP}(O_0) \in \mathbb{R}^{d_{\text{out}}}$, where $\theta = \{\mathbf{P}, \mathbf{S}, \mathbf{C}, \mathbf{S}_k\}$ are the predicted MHR parameters: pose, shape, camera pose and skeleton, respectively. Another set of outputs can be computed from O_{hand} for a pair of MHR hands, which can be merged to the body output to improve the estimation of the hand.

4. Model Training and Inference

Model Training. 3DB is trained with a comprehensive multi-task loss terms, $\mathcal{L}_{\text{train}} = \sum_i \lambda_i \mathcal{L}_i$, where $\mathcal{L}_i$ is a task-specific loss targeting a specific prediction head or anatomical structure. λ_i are hyper-parameters set empirically. In particular, we consider supervision from 2D/3D keypoint locations, MHR parameter regression and hand detection. To stabilize training, certain loss terms (*e.g.*, 3D keypoints) are introduced with a warm-up schedule, gradually increasing their weights over the course of training. We also simulate an interactive setup [17, 46] for training by randomly sampling prompts in multiple rounds per sample. This multi-task, prompt-aware loss design provides strong supervision across all outputs. We describe the losses in the supplementary material.

Full-body Inference. During inference, we use the body decoder output by default, with the option to merge hand decoder output when hands are detected. The benefit of the hand decoder comes from the hand-specific data used during training and the flexibility of a free-moving wrist due to the dedicated prediction head. Specifically, we use the wrist location predicted by the hand decoder and the elbow location from the body decoder to prompt the body decoder to generate a refined full-body pose estimation result. The predicted local MHR parameters are then merged to a full-body configuration following the kinematic tree of the mesh model. Please refer to the supplementary material for more details and qualitative comparisons.

5. Data Annotation and Mesh Fitting

In order to obtain large-scale, high-quality human pose data, we develop a semi-automatic data engine that aims for efficient and divsere image data collection. In addition, we designed a multi-stage annotation pipeline that produces accurate 3D mesh pseudo-ground truth from both in-the-wild single image datasets and a variety of multi-view datasets.

5.1. Data Engine for Data Diversity

Obtaining large-scale human mesh annotations on in-the-wild images can be computationally costly. While it is possible to get a large number of training images from videos, the poses, appearance, imaging conditions, and background might be very similar. In order to increase the diversity of our training dataset, we implemented an automated data engine that selectively routes difficult images for annotation, enabling scalable and efficient dataset curation.

At the core of our data engine is a Vision-Language Model (VLM) driven mining strategy. The VLM identifies images exhibiting challenging scenarios for pose estimation, including occlusion (where the human subject is partially hidden by objects or other people), unusual poses (rare or complex body configurations such as acrobatics or dance), interaction (human-object or human-human activities like holding tools or group actions), extreme scale (subjects appearing at atypical distances from the camera), low visibility (poor lighting, motion blur, or partial visibility), and hand-body coordination (tight coupling of hand and body poses, as in sign language or sports).



Figure 3. Dense (thin) and sparse (thick) keypoints.

Mining rules are automatically updated iteratively based on failure analysis of the current model, allowing the engine to adaptively focus on the most challenging or informative samples. By focusing annotation efforts on the most informative samples, our data engine enables efficient search through tens of millions of images, while maximizing the value and diversity of each annotated image.

5.2. Manual Annotation

Given a set of images selected by the data engine, we use a current version of 3DB to estimate initial 2D joint positions. A team of trained annotators correct the estimated joint locations, if needed. The annotators also assign a per-joint visibility label according to a strict rubric. Joints with substantial occlusion or other factors that would prevent accurate placement (*e.g.*, 50% occlusion, motion blur) are marked as *not visible*.

5.3. Single-Image Mesh Fitting

For each image, we first obtain the initial estimation of MHR parameters from a current version of 3DB's predictions, as well as the 595 dense 2D keypoints predicted from a high-capacity keypoint detector. MHR fitting is then performed via gradient-based refinement of the model parameters, minimizing a composite fitting loss $\mathcal{L}_{\text{fit}} = \sum_j \lambda_j \mathcal{L}_j$, where each $\mathcal{L}_j$ is a task-specific loss including 2D keypoint loss, initialization-anchored regularization and priors. Hyper-parameters λ_j are set via cross-validation.

We apply several loss terms and priors to make the fitting goal: **2D Keypoint Loss** is the L2 distance between projected and detected dense 2D keypoints, to ensure minimal 2D reprojection error. *Initialization-Anchored Regularization* penalizes deviation from the initial prediction by applying L2 losses on both the Momentum Human Rig parameters and their corresponding 3D keypoints, thereby preventing model drift. *Pose and Shape Prior* enforces anatomical plausibility via a learned Gaussian Mixture prior and L2

regularization.

Dense keypoint detector. The configuration of 595 dense keypoints is chosen as it represents the minimal manifold of a human body mesh for capturing diverse body shapes and hand poses. The dense keypoint detector adopts a standard Transformer encoder-decoder architecture [6, 11, 33], with additional sparse keypoint guidance obtained from the manual annotation step to predict accurate 2D dense keypoints from in-the-wild images, as illustrated in Figure 3. We first train the model on 3D datasets (*e.g.*, Goliath and Synthetic), and use it for multi-stage mesh fitting on the in-the-wild datasets (*e.g.*, COCO, AI Challenger, MPII). We then project the MHR mesh to dense keypoints for a second round of dense keypoint detector training and apply this iterative training scheme twice.

5.4. Multi-View Mesh Fitting

While single-view mesh fitting is effective for a large and diverse set of images, the annotation quality tends to be lower fidelity due to the depth ambiguities and natural occlusion. Therefore, we also exploit multi-view mesh fitting on multi-view datasets. Specifically, we extend the single-view pipeline to jointly fit mesh across all frames and camera views, leveraging both spatial and temporal cues.

First, we extract synchronized 2D keypoints for each camera and frame, and then obtain sparse 3D keypoints by triangulation. The mesh model is initialized from these triangulated points and camera parameters, and refined via second-order optimization-based update of the model parameters, minimizing a composite fitting loss, $\mathcal{L}_{\text{multi}} = \sum_k \lambda_k \mathcal{L}_k$, where each $\mathcal{L}_k$ is a task-specific loss including the 2D keypoint loss and the regularization and priors as single-view mesh fitting, together with additional 3D keypoint loss and temporal smoothness. *Temporal Smoothness Loss* encourages estimated pose parameters to temporally smooth, penalizing abrupt changes in motion and promoting realistic temporal dynamics. Optimization alternates between updating camera parameters, shape, skeleton, and pose, with robust keypoint filtering (*e.g.*, robust losses, RANSAC, smoothing). Body shape parameters are optimized jointly across frames.

6. Training Datasets

We train our model on a mix of single-view, multi-view, and synthetic datasets listed in Table 1, covering general body pose, hands, interactions, and "in-the-wild' conditions to ensure the quality, quantity and diversity of training data.

Single-view in-the-wild: We utilize datasets that captures people in unconstrained environments with diverse appearance, pose, and scene conditions. For this, we use AIChallenger [54], MS COCO [25], MPII [1], 3DPW [49], and a subset of SA-1B [17].

Table 1. List of 3DB training datasets. ⋆ denotes the datasets providing samples to train the hand decoder.

Dataset	# Images/Frames	# Subjects	# Views
MPII human pose [1]	5K	5K+	1
MS COCO [25]	24K	24K+	1
3DPW [49]	17K	7	1
AIChallenger [54]	172K	172K+	1
SA-1B [17]	1.65M	1.65M+	1
Ego-Exo4D [10]	1.08M	740	4+
DexYCB [4]	291K	10	8
EgoHumans [15]	272K	50+	15
Harmony4D [16]	250K	24	20
InterHand [29]⋆	1.09M	27	66
Re:Interhand [30]⋆	1.50M	10	170
Goliath [28]⋆	966K	120+	500+
Synthetic⋆	1.63M	–	–

Multi-view consistent: To incorporate geometric consistency for more reliable annotations, we use multi-view data from Ego-Exo4D [10], Harmony4D [16], EgoHumans [15], InterHand2.6M [29], DexYCB [4] and Goliath [28].

High-fidelity synthetic: We use a photorealistic synthetic extension of the Goliath dataset [28]. It provides millions of frames with ground-truth MHR parameters across diverse identities, clothing, and contexts. Synthetic data ensures accurate supervision for human mesh recovery, complementing real-world datasets that prioritize diversity over quality.

Hand datasets: These datasets (marked with ⋆ in Table 1), such as Re:Interhand [30], are used to train both the body and hand decoder. We provide wrist-truncated hand samples to train the hand decoder.

7. Evaluation

We follow prior HMR work and report standard pose and shape evaluation metrics: MPJPE [27], PA-MPJPE [56], PVE [19], and PCK [56]. To evaluate on SMPL-based datasets, a MHR mesh is mapped to the SMPL mesh format. We present results with two variants of the model; 3DB-H leverages the commonly used ViT-H (632M) backbone, and 3DB-DINOv3 uses the recent DINOv3 (840M) [45] encoder. We resize the input to 512×512 for the image encoder and use an off-the-shelf field-of-view (FOV) estimator (*e.g.*, MoGe-2 [50]) to provide camera intrinsics for model inference.

7.1. Evaluating Performance on Common Datasets

We first evaluate 3DB on five standard benchmark datasets in Table 2, comparing with a wide variety of state-of-the-art (SoTA) mesh recovery methods. 3DB outperforms all other single-image methods and is even competitive with video-based approaches that additionally leverage temporal information.

In particular, our model achieves superior results in the EMDB and RICH datasets, which are *out-of-domain* (*i.e.*,

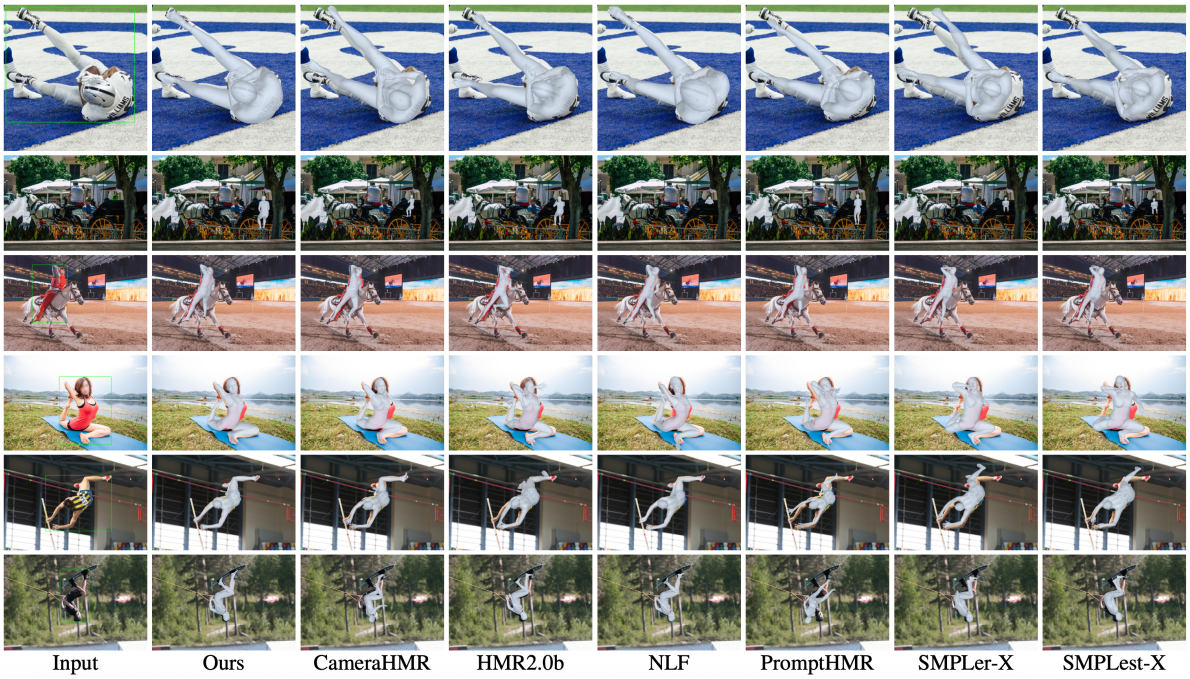


Figure 4. Qualitative comparison of 3DB against state-of-the-art HMR methods.

Table 2. Comparison on five common benchmarks. The best results are highlighted in bold, while the second-best results are underlined. Results evaluated using publicly released checkpoint denoted by †. Models trained using RICH denoted by *.

		3DPW (14)			EMDB (24)			RICH (24)			COCO	LSPET
	Models	PA-MPJPE ↓	MPJPE ↓	PVE ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓	PCK@0.05 ↑	PCK@0.05 ↑
IMAGE	HMR2.0b [9]	54.3	81.3	93.1	79.2	118.5	140.6	48.1†	96.0†	110.9†	86.1	53.3
IMAGE	CameraHMR [33]	35.1	56.0	65.9	43.3	70.3	81.7	34.0	55.7	64.4	80.5†	49.1†
IMAGE	PromptHMR [53]	36.1	58.7	69.4	41.0	71.7	84.5	37.3	56.6	65.5	79.2†	55.6†
IMAGE	SMPLerX-H [3]	46.6†	76.7†	91.8†	64.5†	92.7†	112.0†	37.4†	62.5†	69.5†	–	–
IMAGE	NLF-L+fit* [43]	33.6	54.9	63.7	40.9	68.4	80.6	28.7†	51.0†	58.2†	74.9†	54.9†
VIDEO	WHAM [44]	35.9	57.8	68.7	50.4	79.7	94.4	–	–	–	–	–
VIDEO	TRAM [52]	35.6	59.3	69.6	45.7	74.4	86.6	–	–	–	–	–
VIDEO	GENMO [20]	34.6	**53.9**	65.8	42.5	73.0	84.8	39.1	66.8	75.4	–	–
	3DB-H (Ours)	**33.2**	54.8	64.1	38.5	62.9	74.3	31.9	55.0	61.7	**86.8**	**68.9**
	3DB-DINOv3 (Ours)	33.8	54.8	**63.6**	**38.2**	**61.7**	**72.5**	**30.9**	**53.7**	**60.3**	86.5	67.8

not included in the training set), indicating better generalization than previous SoTA methods. 3DB exceeds the second best model, NLF, on all datasets in terms of 3D metrics except for RICH which dataset NLF uses in training while our model does not. 3DB is also state-of-the-art on PCK for 2D evaluation on the COCO and LSPET datasets, demonstrating strong 2D alignment.

7.2. Evaluating Performance on New Datasets

Throughout our experiments, we found that mesh recovery models are particularly fragile in out-of-domain settings due to camera, appearance, and pose differences. To understand how methods perform on new, unseen data distributions, we additionally evaluate on five new datasets (38.6K images) in Table 3. The five new datasets include (1) Ego-Exo4D [10], (2) Harmony4D [16], (3) Goliath [28], (4) in-house synthetic data and (5) SA1B-Hard. Ego-Exo4D captures humans in diverse, skilled activities, divided into physical (EE4D-Phys) and procedural (EE4D-Proc) domains. Harmony4D focuses on close multi-human interaction in dynamic sports settings. Goliath offers diverse motions in a precise, studio environment. The synthetic dataset consists of single-human images with diverse camera angles and parameters. SA1B-Hard is a subset of 2.6K images extracted from SA1B using our data engine. Together, these five new datasets present a challenging new testbed for mesh recovery methods.

As it is difficult to compare methods using the exact same training data and methodology due to prohibitive data usage licenses, unclear descriptions of training data, and

Table 3. Comparison on five new benchmark datasets. The best results are highlighted in bold, while the second-best results are underlined. MPJPE is computed on 24 SMPL keypoints.

	EE4D-Phy		EE4D-Proc		Harmony4D		Goliath		Synthetic		SA1B-Hard
Models	PVE ↓	MPJPE ↓	PVE ↓	MPJPE ↓	PVE ↓	MPJPE ↓	PVE ↓	MPJPE ↓	PVE ↓	MPJPE ↓	Avg-PCK ↑
CameraHMR [33]	<u>71.1</u>	<u>58.8</u>	<u>70.3</u>	<u>60.2</u>	<u>84.6</u>	<u>70.8</u>	66.7	<u>54.5</u>	102.8	87.2	63.0
PromptHMR [53]	74.6	63.4	72.0	62.6	91.9	78.0	67.2	56.5	<u>92.7</u>	<u>80.7</u>	59.0
NLF [43]	75.9	68.5	85.4	77.7	97.3	84.9	<u>66.5</u>	58.0	97.6	86.5	<u>66.5</u>
3DB-H Leave-one-out (Ours)	**49.7**	**44.3**	**52.9**	**47.4**	**63.5**	**54.0**	**54.2**	**46.5**	**85.6**	**75.5**	**73.1**
3DB-H Full dataset (Ours)	37.0	31.6	41.9	36.3	41.0	33.9	34.5	28.8	55.2	47.2	76.6

Table 4. Comparison on Freihand for hand pose estimation. Methods using Freihand for training are denoted by †.

Method	PA-MPVPE ↓	PA-MPJPE ↓	F@5 ↑	F@15 ↑
LookMa [11]	8.1	8.6	0.653	-
METRO [24]†	6.3	6.5	0.731	0.984
HaMeR [35]†	5.7	6.0	0.785	0.990
MaskHand [42]†	5.4	5.5	0.801	0.991
WiLoR [38]†	5.1	5.5	0.825	0.993
3DB-H (Ours)	6.3	5.5	0.735	0.988
3DB-DINOv3 (Ours)	6.2	5.5	0.737	0.988

lack of training code (CameraHMR, PromptHMR, and NLF are trained on 6, 9, and 48 datasets, respectively), we test the generalization ability of 3DB by using a leave-one-out training procedure. This ensures a fair comparison with prior work which have also not seen these datasets. To serve as an in-domain, upper bound comparison, we also show the performance of 3DB when trained on the *full dataset* (*i.e.*, training data is also sampled from these new datasets). For both the baselines and our model, we use ground truth camera intrinsics for model inference on 3D datasets, except for SA1B-Hard which we use FOV estimated by MoGe-2 [50].

We present the results in Table 3. Despite being trained on a large number of datasets, we find that prior work still struggle with these five domains, incurring a significant drop in performance. In contrast, our leave-one-out model shows strong generalization, owing to our more diverse data distribution and stronger training framework. Interestingly, we notice that existing methods constantly trade places for second across different datasets, reflecting strong dataset-specific biases. This indicates that each baseline overfit to a narrow slice of the underlying data distribution.

7.3. Evaluating Hand Pose Estimation Performance

One significant characteristic of 3DB is its strong performance in estimating hand shape and pose. Previous full-body human pose estimation methods [2, 3, 23] revealed a notable gap in hand pose accuracy compared to *hand-only* pose estimation methods [35, 38]. This performance gap arises from two main factors. First, hand-only methods can leverage large-scale datasets of hand poses, whereas full-body methods cannot utilize these datasets because of the absence of full-body images and annotations. Second, a free-moving wrist allows hand pose models to more easily fit finger poses with 2D and 3D alignment, while for full-body methods, wrist rotation and position are highly constrained by the body's pose and position. Despite these challenges, 3DB demonstrates strong hand pose accuracy. 3DB benefits from the flexible model training design that incorporates both hand and body data and the hand decoder. Additionally, being promptable, 3DB provides a natural mechanism to align the wrists of the body prediction with those of the hands. We evaluate 3DB's hand estimation on the representative FreiHand [57] benchmark in Table 4. For fair comparison against hand-only models, we use the output from our hand decoder for evaluation. Despite not training on the Freihand dataset, which gives a strong in-domain boost, 3DB's hand pose estimation accuracy is already comparable to SoTA hand pose estimation methods that include Freihand alongside many other hand-centric datasets.

7.4. Evaluating 2D Categorical Performance

To better understand the strengths and weaknesses of models on a variety of image types, we compare the performance across our 24 categories defined over SA1B-Hard [17]. Our proposed evaluation set is designed to capture a broad spectrum of human appearance and activity in images, ensuring robust evaluation across real-world scenarios. It consists of 24 total categories, which are organized under several high-level groups: Body Shape, Camera View, Hand, Multi-person, Pose and Visibility.

We use the PCK (Percentage of Correct Keypoints) metric for 17 body keypoints and 6 feet keypoints. Results are reported using Avg-PCK, which is PCK averaged over a range of thresholds (*i.e.* 0.01, 0.025, 0.05, 0.075, 0.1 of the human bounding box size). Results in Table 5 show that 3DB outperforms all baselines on all categories. Qualitative examples are given in Figure 4.

One notable significance is for categories of *Visibility - Truncation* where the model shows significant advantages than CameraHMR or PromptHMR. Essentially, 3DB has learned a much stronger pose prior when dealing with body truncation in images. Other rows with the large improve-

Table 5. 2D categorical performance analysis on the SA-1B Hard dataset.

	CameraHMR [33]		PromptHMR [53]		3DB	
	APCK(body)	APCK(feet)	APCK(body)	APCK(feet)	APCK(body)	APCK(feet)
Body_shape - In-the-wild	87.64	78.56	85.73	77.87	**90.76**	**92.12**
Camera_view - Back or side view	59.69	46.64	61.92	47.74	**76.27**	**66.81**
Camera_view - Bottom-up view	55.18	34.84	46.56	29.25	**69.62**	**55.35**
Camera_view - Others	51.48	33.80	54.39	38.55	**76.62**	**71.52**
Camera_view - Overhead view	55.08	39.46	43.65	24.63	**73.33**	**66.94**
Hand - Crossed or overlapped fingers	73.20	62.85	72.48	62.43	**81.36**	**84.04**
Hand - Holding objects	76.73	72.11	73.57	68.92	**83.40**	**85.92**
Hand - Self-occluded hands	73.22	58.06	72.43	56.19	**80.07**	**80.82**
Multi_people - Contact or interaction	63.23	51.65	61.77	47.60	**74.81**	**69.92**
Multi_people - Overlapped	53.11	41.88	57.17	41.43	**70.82**	**64.71**
Pose - Contortion or bending	47.08	32.78	42.61	20.98	**65.20**	**53.04**
Pose - Crossed legs	63.95	32.24	56.15	27.35	**76.40**	**58.80**
Pose - Inverted body	46.12	30.01	39.83	24.64	**78.18**	**72.19**
Pose - Leg or arm splits	57.51	31.43	54.76	33.11	**83.69**	**72.49**
Pose - Lotus pose	63.19	14.38	54.85	12.87	**74.53**	**57.97**
Pose - Lying down	51.29	35.88	44.59	26.88	**71.35**	**66.53**
Pose - Sitting on or riding	79.66	71.65	70.15	61.16	**84.85**	**81.51**
Pose - Sports or athletic activities	78.93	69.34	73.62	60.37	**85.10**	**82.80**
Pose - Squatting or crouching or kneeling	62.74	41.47	54.41	33.84	**72.85**	**61.85**
Visibility - Occlusion (foot cues)	62.93	26.83	58.00	30.81	**75.43**	**54.74**
Visibility - Occlusion (hand cues)	61.01	53.89	58.55	51.13	**76.04**	**72.01**
Visibility - Truncation (lower-body truncated)	39.27	-	46.50	-	**61.95**	-
Visibility - Truncation (others)	79.18	74.82	77.06	74.99	**84.23**	**86.72**
Visibility - Truncation (upper-body truncated)	62.37	54.90	56.01	49.28	**64.49**	**70.99**

ments are *Pose - Inverted body* and *Pose - Leg or arm splits*. We largely attribute these improvements to the increased distribution of hard poses selected by the data engine.

7.5. Human Preference Study

We conducted a large-scale user preference study to evaluate the perceptual quality of human reconstructions produced by 3DB compared with existing approaches on the SA1B-Hard dataset. While quantitative metrics capture geometric and numeric accuracy, they do not always align with the human perception accuracy.

We designed six independent pairwise comparison studies, each comparing 3DB against one baseline method: HMR2.0b [9], CameraHMR [33], NLF [43], PromptHMR [53], SMPLer-X [3], and SMPLest-X [55]. The study encompassed $7,800$ unique participants ($1,300$ unique per comparison) resulting in over $20,000$ total responses. Each participant was presented with a video stimuli. The left and right sides of the video displayed reconstructions from the two methods, and a video transition effect as used to fade-in the reconstruction result over the image. Participants were instructed to choose which 3D reconstruction better matched the original image by answering: *"Which 3D model of the person better matches the original image, left or right?"*. We quantify results using win rate and vote share. Win rate is the percentage of stimuli for which 3DB received more votes than the baseline. As summarized in Figure 5, 3DB consistently outperforms all baselines. Focusing on the strongest baseline, NLF, 3DB achieves a win rate of 83.8%.

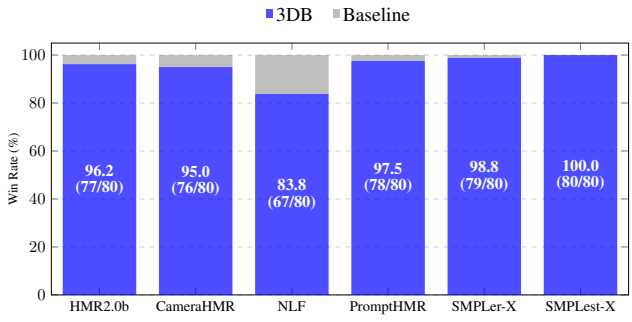


Figure 5. Comparison of 3DB win rate against baselines. Win rate (%) and number of wins out of 80.

8. Conclusion

We have presented 3DB, a robust HMR model for body and hands. Our approach leverages the Momentum Human Rig parametric body model, employs a flexible encoder–decoder architecture, and supports optional prompts such as 2D keypoints or masks to guide inference. A central advance of our work is in the supervision pipeline. Instead of relying on noisy monocular pseudo-ground-truth, we leverage multi-view capture systems, synthetic sources, and a scalable data engine that actively mines and annotates challenging samples. This strategy yields cleaner and more diverse training signals, supporting generalization beyond curated benchmarks. At the same time, 3DB employs a separate hand decoder to enhance the hand pose estimation with hand crops as input which makes it comparable to SoTA hand pose estimation methods.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1, 5

[2] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 2, 7

[3] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023. 2, 6, 7, 8

[4] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 1, 5

[5] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40. Springer, 2020. 2

[6] Hanz Cuevas-Velasquez, Anastasios Yiannakidis, Soyong Shin, Giorgio Becherini, Markus Höschle, Joachim Tesch, Taylor Obersat, Tsvetelina Alexiadis, Eni Halilaj, and Michael J Black. Mamma: Markerless & automatic multi-person motion action capture. *arXiv preprint arXiv:2506.13040*, 2025. 5

[7] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1323–1333, 2024. 1, 2

[8] Aaron Ferguson, Ahmed AA Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, et al. Mhr: Momentum human rig. *arXiv preprint arXiv:2511.15586*, 2025. 2

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 1, 2, 6, 8

[10] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 5, 6

[11] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look ma, no markers: holistic performance capture without the hassle. *ACM Transactions on Graphics (TOG)*, 43(6), 2024. 5, 7

[12] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013. 1

[13] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 2017. 1

[14] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2

[15] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An egocentric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19807–19819, 2023. 5

[16] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4d: A video dataset for in-the-wild close human interactions. *Advances in Neural Information Processing Systems*, 37:107270–107285, 2024. 2, 5, 6

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3, 4, 5, 7

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 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*, pages 2252–2261, 2019. 2

[19] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 5

[20] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11766–11776, 2025. 6

[21] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2

[22] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 2

[23] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 7

[24] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. 7

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

[27] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 5

[28] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, et al. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. *Advances in Neural Information Processing Systems*, 37:83008–83023, 2024. 2, 5, 6

[29] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*, pages 548–564. Springer, 2020. 2, 5

[30] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A dataset of relighted 3d interacting hands. *Advances in Neural Information Processing Systems*, 36:17689–17701, 2023. 5

[31] Jinhyung Park, Javier Romero, Shunsuke Saito, Fabian Prada, Takaaki Shiratori, Yichen Xu, Federica Bogo, Shoou-I Yu, Kris Kitani, and Rawal Khirodkar. Atlas: Decoupling skeletal and shape parameters for expressive parametric human modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6508–6518, 2025. 2

[32] Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos. *arXiv preprint arXiv:2211.13225*, 2022. 1

[33] Priyanka Patel and Michael J Black. Camerahmr: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pages 1562–1571. IEEE, 2025. 1, 2, 5, 6, 7, 8

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 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*, pages 10975–10985, 2019. 2

[35] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 2, 7

[36] Owen Pearl, Soyong Shin, Ashwin Godura, Sarah Bergbreiter, and Eni Halilaj. Fusion of video and inertial sensing data via dynamic optimization of a biomechanical model. *Journal of biomechanics*, 155:111617, 2023. 1

[37] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions On Graphics (TOG)*, 37(6):1–14, 2018. 1

[38] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12242–12254, 2025. 2, 7

[39] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[40] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2

[41] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021. 2

[42] Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Mayur Jagdishbhai Patel, Hongfei Xue, Ahmed Helmy, Srijan Das, and Pu Wang. Maskhand: Generative masked modeling for robust hand mesh reconstruction in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8372–8383, 2025. 7

[43] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. *Advances in Neural Information Processing Systems*, 37: 140032–140065, 2024. 6, 7, 8

[44] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 6

[45] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 5

[46] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 4

[47] Zhihao Sun, Tong Wu, Ruirui Tu, Daoguo Dong, and Zuxuan Wu. Unihand: A unified model for diverse controlled 4d hand motion modeling. *arXiv preprint arXiv:2602.21631*, 2026. 2

[48] Vasileios Vasilopoulos, Georgios Pavlakos, Sean L Bowman, J Diego Caporale, Kostas Daniilidis, George J Pappas, and Daniel E Koditschek. Reactive semantic planning in unexplored semantic environments using deep perceptual feedback. *IEEE Robotics and Automation Letters*, 5(3):4455–4462, 2020. 1

[49] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 1, 5

[50] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 5, 7

[51] Shengze Wang, Jiefeng Li, Tianye Li, Ye Yuan, Henry Fuchs, Koki Nagano, Shalini De Mello, and Michael Stengel. Blade: Single-view body mesh estimation through accurate depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21991–22000, 2025. 1, 2

[52] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487. Springer, 2024. 6

[53] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1148–1159, 2025. 1, 2, 6, 7, 8

[54] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *International Conference on Multimedia and Expo (ICME)*, pages 1480–1485. IEEE, 2019. 5

[55] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, et al. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2, 8

[56] Jianfeng Zhang, Xuecheng Nie, and Jiashi Feng. Inference stage optimization for cross-scenario 3d human pose estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:2408–2419, 2020. 5

[57] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 7