All-Age Human Mesh Recovery

Laura Bravo-Sánchez[1,2] Matthieu Armando[1] Romain Brégier[1] Grégory Rogez[1]
Serena Yeung-Levy[2] Fabien Baradel [1]

[1]NAVER LABS Europe, [2]Stanford University,

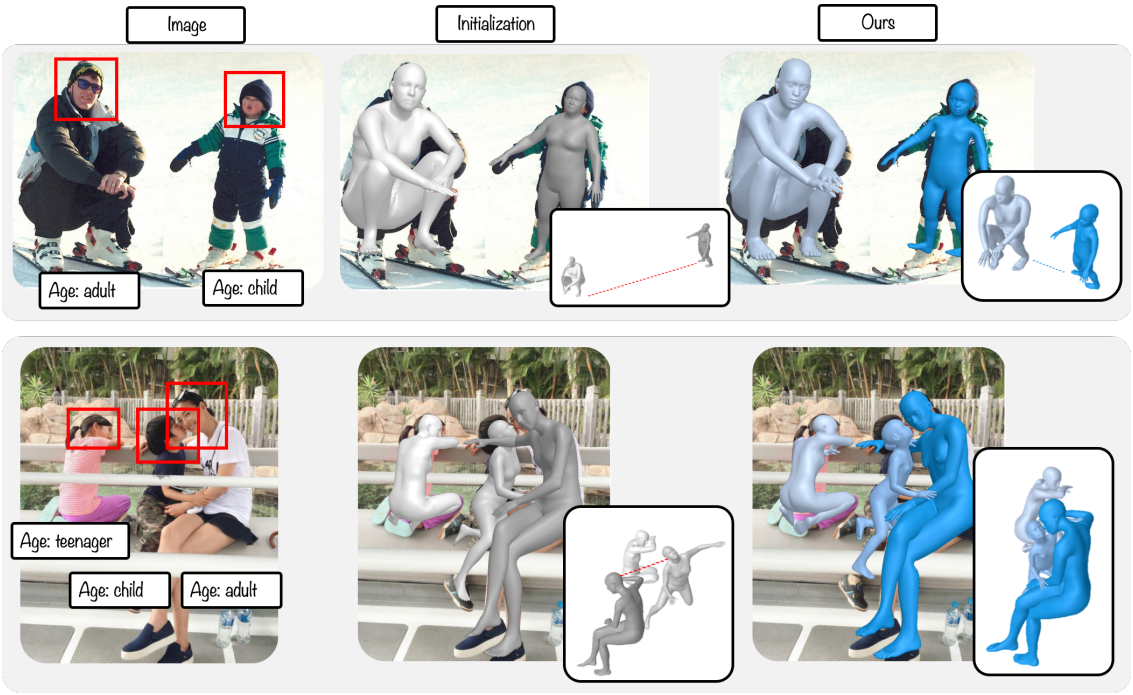

Figure 1. **Anny-Fit** recovers multi-person 3D human meshes of all ages directly in camera space. By integrating expert semantic, depth, keypoint, and segmentation cues, it improves all-age HMR and enables zero-shot adaptation of adult-only models.

Abstract

*Recovering 3D human pose and shape from a single image remains a cornerstone of human-centric vision, yet most methods assume adult subjects and optimize each person independently. These assumptions fail in real-world, all-age scenes, where body proportions and depth must be resolved jointly. We introduce **Anny-Fit**, a multi-person, camera-space optimization framework for all-age 3D human mesh recovery (HMR). Unlike existing per-person fitting methods, Anny-Fit jointly optimizes all individuals directly in the camera coordinate system, enforcing global spatial consistency. At the core of our approach is the use of multiple forms of expert knowledge—including metric depth maps, instance segmentation, 2D keypoints, and, VLM-derived semantic attributes such as age and gender—each obtained from dedicated off-the-shelf networks. These complementary signals jointly guide the optimization, constraining the depth-scale ambiguity characteristic of all-age scenes. Across diverse datasets, Anny-Fit consistently improves 2D reprojection accuracy (+13 to 16), relative depth ordering (+6 to 7), 3D estimation error (-9 to -29) and shape estimation (+25 to +82), producing more coherent scenes. Finally, we show that VLM-based semantic knowledge can be distilled into an HMR model via the pseudo-ground-truth annotations produced by Anny-Fit on training data, enabling it to learn semantically meaningful shape parameters while improving HMR performance. Our approach bridges adult-only and all-age modeling by enabling zero-shot adaptation of adult-trained HMR pipelines to the full age spectrum without retraining.*

1. Introduction

Reconstructing 3D human pose and shape from a single monocular image (HMR) is a fundamental problem in human-centric scene understanding, pivotal for applications such as robotic navigation and embodied AI. The central challenge of this monocular task is resolving the inherent depth ambiguity: a person's apparent size in an image is a joint function of both their physical stature and their distance from the camera (Fig.2). Much prior work circumvents this ambiguity by assuming all subjects are adults [54, 66], allowing apparent size to serve as a reliable depth cue. However, this assumption breaks down in realistic, all-age settings, where the problem becomes significantly more under-constrained: a small silhouette could correspond to either a distant adult or a nearby child. Consequently, apparent size alone is insufficient, necessitating the joint estimation of depth and body shape.

Because monocular HMR is under-specified, most approaches leverage external information to constrain the solution space. Such information may come from the scene (e.g., camera parameters [2, 64, 66], person location [17, 56], depth cues [52], or contact maps [5, 35, 36, 53]) or from the humans themselves (e.g., 2D keypoints [4, 25], dense points [38, 55, 63, 67] or shape cues [6, 9]). Among these, we argue that the body shape cues are particularly critical for addressing the all-age ambiguity. Figure 2 illustrates how estimating whether a subject corresponds to an adult or a child re-constrains the depth-shape ambiguity, facilitating reliable 3D recovery.

While shape cues and an expressive body model help resolve ambiguity at the individual level, they do not by themselves guarantee *scene-level* coherence in multi-person settings. Fitting each person independently to their 2D evidence often leads to inconsistent relative depths that break the scene's spatial layout. We posit that recovering a valid 3D layout requires enforcing relational depth consistency across all subjects.

To address these challenges, we introduce **Anny-Fit**, a flexible optimization framework that integrates both scene- and person-level constraints to refine initial estimates. Importantly, these constraints extend beyond human annotation: Anny-Fit leverages specialized models (such as detectors, depth estimators, and 2D keypoint regressors) together with generalist Vision-Language Models (VLMs) whose high-level semantic predictions (e.g., estimated age and gender) can be translated into meaningful shape parameters, which HMR models struggle to infer reliably. This combination enables Anny-Fit to operate either fully automatically or in a semi-supervised fashion.

A significant challenge to adapting HMR for all-age is the availability of body models that reflect different ages. While prior work [39, 52] introduced SMPL-A to interpolate between infant and adult extremes, we adopt the more recent Anny [7] model, which offers two key advantages for our purposes. First, Anny provides, with a single model, a continuous representation of shape variation across the full human lifespan from infants to seniors. This broader representational coverage enables consistent reasoning about shape and depth in complex multi-person all-age scenes. Second, Anny's shape space is parameterized by semantic attributes (such as age, gender, height, and weight) that align naturally with observable image cues. We exploit these two advantages with our key insight: mapping these continuous semantic attributes to discrete categories already understood by general-purpose 2D vision models. Based on this, we propose to repurpose VLMs as a training-free approach to shape estimation.

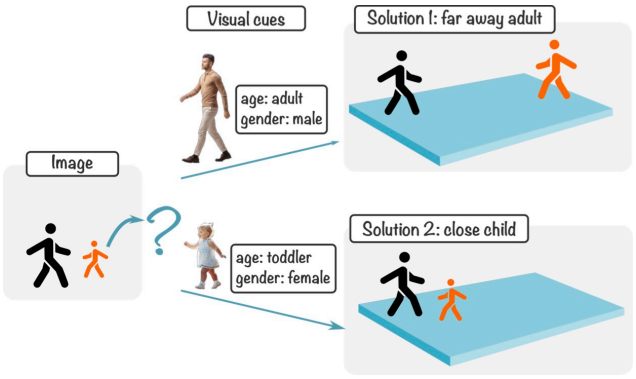

Figure 2. **Depth-scale ambiguity.** Unlike adult-only settings where body size reliably indicates depth, the all-age setting generalizes the problem such that size alone cannot distinguish depth: identical 2D reprojections can correspond to either distant adults or nearby children. We leverage visual cues to infer shape and re-constrain the problem.

We build Anny-Fit on an optimization-based paradigm, as such methods have proven vital in HMR. Optimization methods serve dually as a post-processing step to align regression-based predictions with image evidence [55, 63] and as a generator of pseudo-ground-truth [36, 38, 58] to scale training beyond small curated in-the-wild datasets and lab-controlled captures. Building on this foundation, Anny-Fit combines semantic shape initialization, joint camera-space optimization, and multi-source expert cues, enabling fully automatic reconstruction and the large-scale creation of high-quality pseudo-ground-truth for all-age, multi-person scenes.

In summary, our main contributions are:

- A new formulation of all-age human mesh recovery. We highlight the limitations of adult-only HMR and characterize the depth–shape ambiguity that arises in real-world all-age scenes.
- We introduce Anny-Fit, a camera-space HMR optimization framework that jointly reconstructs multiple individuals while enforcing relational depth consistency.
- We propose a principled fusion of expert cues to guide

all-age initialization and optimization. Our results show that Anny-Fit can improve existing all-age models and enable zero-shot adaptation of adult-only HMR models to the full age spectrum.
- We demonstrate that Anny-Fit can generate high-quality semantic pseudo-ground truth at scale, facilitating downstream HMR models to learn semantically meaningful shape parameters while increasing accuracy.

2. Related work

Multi-person HMR. Human Mesh Recovery (HMR) aims to estimate full 3D human bodies from monocular images [23]. Parametric methods [7, 31, 40, 59] dominate this task by regressing body model parameters. While early works focused on single-person settings [23, 25], recent advances address multi-person HMR [2, 6, 49, 50], where all people must be localized and recovered coherently in 3D.

Two broad paradigms exists: regression and optimization. Regression-based methods [2, 49–52, 56] directly predict human meshes from image features, offering efficiency but limited generalization to unseen demographics or complex occlusions. Optimization-based approaches [22, 25, 38] instead refine predicted mesh parameters to better match image cues, providing higher fidelity but depending heavily on initialization. Both rely on strong priors, including pose regularization [10, 32, 40], camera intrinsics [2], human height [22], or additional cues such as segmentation and text [56]. An additional challenge in multi-person scenes is reasoning about spatial relationships. Prior works improve relative depth via camera-space formulations [2, 27, 38, 64], depth supervision [22, 52], or explicit interaction modeling [16, 20, 24, 37]. In contrast, we propose a simple depth-ordering loss leveraging recent advances in monocular depth estimation to refine global 3D placement and ensure consistent camera-space positioning across all subjects given per-person shape priors.

Modeling age in HMR. Most existing HMR approaches have been designed and evaluated primarily for adults. Beyond modeling limitations, this strong adult bias also reflects data scarcity: large-scale annotated datasets containing children are extremely limited due to (i) the rarity of publicly available child imagery on the web, (ii) strict privacy and consent constraints for minors, and (iii) the ethical sensitivity surrounding the collection and release of children's images. As a result, current training corpora overwhelmingly depict adults, making it difficult for standard HMR pipelines to generalize across the full age spectrum.

Early work by Hesse *et al*. [18, 19] introduced SMIL, a parametric model of infants (2–4 months old) derived from 3D scans. To extend age coverage, AGORA [39] employed SMPL-XA, a piecewise interpolation between SMPL-X [31, 40] and SMIL to generate synthetic child bodies, but the resulting shapes were often inconsistent due to the discontinuity between adult and infant morphologies. BEV [52] was the first to tackle all-age estimation from in-the-wild data using weak supervision from age categories, depth layers, and 2D keypoints. HARMONI [57] further explored joint modeling of SMPL and SMIL for analyzing longitudinal child growth.

Recently, Anny [7] introduced a parametric model covering the full human lifespan, making it particularly suitable for modeling age in HMR. However, as a new body model, Anny currently lacks the large-scale training data and broad ecosystem support available for more established models like SMPL [31] or SMPL-X [40]. This limits the effectiveness of purely regression-based learning approaches built on top of it. To address this, we propose an optimization-based method that explicitly incorporates VLM-derived age cues to guide the fitting process, enabling reliable all-age 3D recovery in a low-data regime.

Language-guided human shape modeling. A complementary direction in HMR focuses on modeling body shape using semantic information rather than purely geometric cues. Early works model adult shape variation based on size and height [54], while methods such as STRAPS [46], BodyTalk [48], and SHAPY [9] enrich shape estimation by leveraging textual human attributes (e.g., height, weight, body type). SHAPY is particularly relevant to our setting as it demonstrates that semantic descriptions can guide fine-grained shape prediction from single images.

Datasets such as SSP-3D [46] provide valuable 3D supervision for shape-aware prediction, yet remain limited in scale and in body-shape diversity—especially for subjects outside the adult age range—highlighting the need for alternative sources of semantic conditioning. On the other hand, facial modeling research has explored estimating human attributes such as age and other semantics directly from single images [44, 47]. Parallel progress in text–image alignment has shown that language supervision provides a rich representation of human pose and shape attributes [44, 47] to joint embedding models that reason over visual and textual descriptions [11–13, 15]. Inspired by this, we depart from training a dedicated body-shape regressor and instead leverage foundation Vision–Language Models (VLMs) [1, 29, 30] as general-purpose estimators of semantic human characteristics. By interfacing VLM predictions with the semantic shape space of the Anny body model [7], we translate text-derived cues directly into the shape space during optimization. This enables flexible, data-free semantic conditioning across diverse subjects without requiring large-scale shape-labeled datasets, and provides a natural bridge between text-driven supervision and 3D body shape estimation.

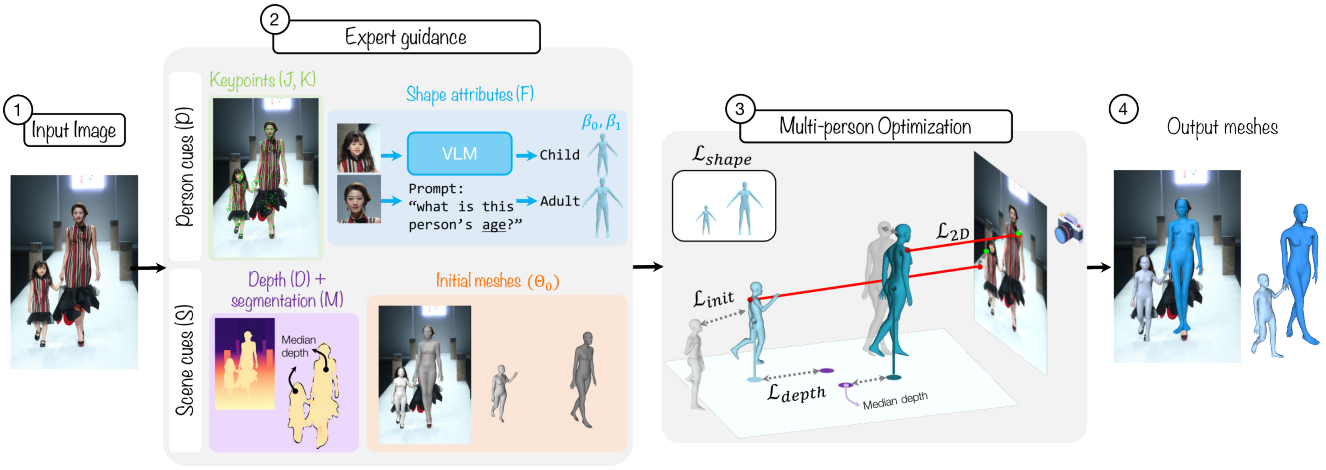

Figure 3. **Overview of Anny-Fit**. We refine initial human meshes estimated using an HMR network through iterative optimization. Anny-Fit leverages pre-computed cues from expert vision models to guide fitting. To mitigate degenerate depth solutions that satisfy 2D reprojection losses, we incorporate shape attribute estimation and an explicit multi-person depth loss.

3. Method

3.1. Problem formulation

Our goal is to recover 3D posed human meshes of individuals of all ages from monocular RGB images. We represent each individual in the scene with the 3D parametric all-age human body model Anny [7], chosen for its explicit parameterization of age-dependent shape variation. The model parameters for person i are $\Theta^i = \{\beta^i, \phi^i, \tau^i, \theta^i\}$. Where $\beta^i \in \mathbb{R}^{10}$ denotes the shape parameters, $\phi^i \in \mathbb{R}^3$ the root orientation, $\tau^i \in \mathbb{R}^3$ the root translation, and $\theta^i \in \mathbb{R}^{163}$ the articulated pose. Given an input image I containing N individuals of varying ages, our objective is to estimate a set of parameters $\Theta_{\text{init}} = \{\Theta^1, \ldots, \Theta^N\}$ from image-derived expert knowledge, and refine them through optimization to obtain image-aligned parameters Θ_{final}.

3.2. Anny-Fit

Figure 3 provides an overview of our approach for multi-person, all-age human mesh recovery. The core idea is to treat the task as an expert-guided optimization, where auxiliary cues from expert models provide strong priors helping to disambiguate the mesh recovery problem. We organize these cues into person-level $\mathcal{P}$ and scene-level constraints $\mathcal{S}$, which provide weak supervision signals to guide the fitting process. In particular, the person-level cues $\mathcal{P} = \{J, F, K\}$, consist of 2D joints location J, shape attributes estimates F and dense 2D keypoints locations K. While the scene-level cues $\mathcal{S} = \{D, M, \Theta_{t-1}\}$, stem from metric depth maps (D) and instance segmentation (M) predictions, as well as the previous optimization state Θ_{t-1}, used as a regularizer to prevent excessive drift during the optimization. These cues are integrated into the overall objective function, detailed below.

Person cues. Central to age estimation is the shape-depth ambiguity problem shown in Figure 2. Without a close initial shape estimate or a perfect body pose and positioning, the optimization fixes the easiest one of these often biasing the solution to an unwanted local minima. This can happen by having a model that lacks understanding of the visual representation of a person's shape or has trouble placing a person in a scene. We address this challenge by building on the interpretable semantic shape space of the Anny model, for which each dimension of β corresponds to non-independent physical attribute (age, gender, weight, height, muscle, etc.) and utilizing external knowledge derived from human annotation or from a foundation model.

We show how a generalist model like a VLM that has visual appearance information encoded can be queried to initialize the shape of each person. In particular, we use [1] as our VLM to constrain the shape optimization two-fold. First, we obtain an image-aligned estimate of β, denoted F, which is a mapping to the normalized Anny space of a predicted category label. Second, we use F to constrain deviations throughout the optimization via a loss term $\mathcal{L}_{shape} = MSE(\beta, F)$. We query for F by cropping each person's head using J when available, and otherwise by taking the detection bounding box (see Supp. Mat. for details on prompts).

We note that F need not be exact but close proxy to simplify the optimization. We avoid directly regressing continuous attributes (e.g., chronological age) from the VLM, as this is notoriously difficult. Prior work [3, 15], has shown the limited effectiveness of direct regression largely because it requires vast training data to overcome the limitations of tokenizers. Furthermore, such continuous attributes are often conceptually ambiguous. For instance, chronological age is an unreliable proxy for visual appearance, as two individuals of the same age can have vastly different biological ages and, consequently, different body shapes. Given

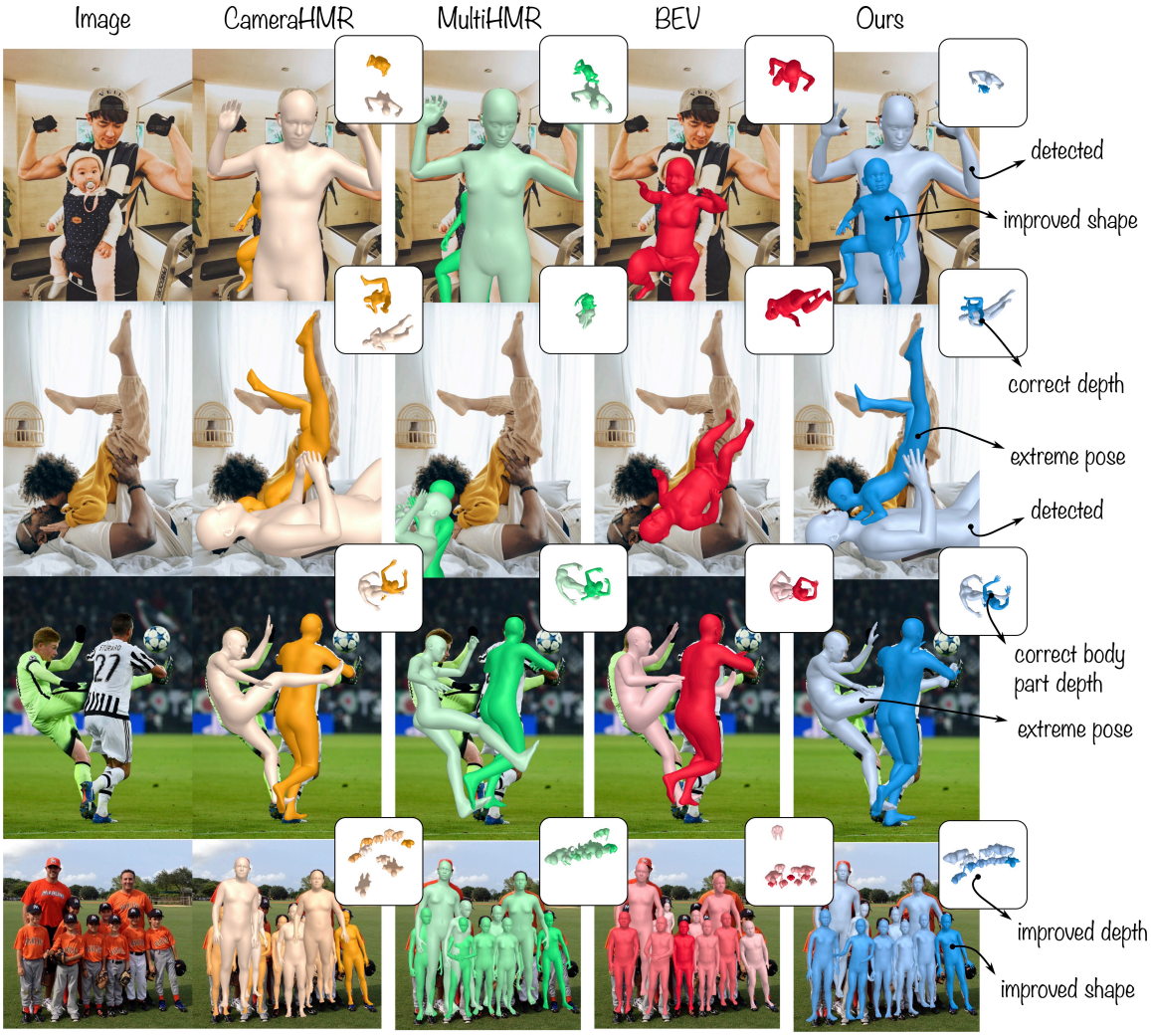

Figure 4. **Qualitative results**, in front and top view. Our method, Anny-Fit, exploits the advantages of SOTA models and generalizes them to the all-age setting. Compared to BEV, this translates to improvements in detection, depth ordering, shape and pose estimation.

these challenges, our key insight is to reformulate shape cue estimation as a categorization task over each dimension of β (which has been shown to be more effective for VLM prediction [14, 65]). We select anchor points with semantic meaning and remaping each prediction in the normalized parameter space of β. In our experiments, we focus on two primary axes of variation—age and gender—while noting that this process can be naturally extended to capture a wider range of body shapes. Namely for age we select six anchor points covering the human life-span, with more detailing on early ages where shape has rapid change: 'baby', 'toddler', 'child', 'teenager', 'adult', and 'senior'. For gender we select 3 anchors: 'male', 'neutral' and, 'female'. For all attribute types we set a fallback 'unknown' as the center of the range.

In addition, we rely on J and K, which are both sets of 2D points p_j to guide optimization to the image. We assume that there is a known direct one-to-one correspondence between each p_j and some 3D point q_j attached to the body model. Our losses are then $\mathcal{L}_{2D} = \mathcal{L}_{dense} = \frac{1}{|V|}\sum_{j\in V}\rho(c_j\|\hat{p}_j - p_j\|_2, \sigma)$, where $\hat{p}_j = \Pi(q_j)$ is the reprojection of q_j on the image plane, using known or estimated camera intrinsics. $c_j \in [0, 1]$ is the confidence of point j, V is the set of points (J or K) and $\rho(x, \sigma) = \frac{\sigma^2 x^2}{\sigma^2 + x^2}$ is the Geman-McClure robust error function to handle outlier values. We obtain J and K from estimators [60] and [38] respectively.

Multi-person optimization. To obtain Θ_{final}, we perform a full-scene, multi-stage optimization that refines the set of meshes Θ_t jointly across all individuals. Unlike prior approaches that optimize per-person crops and remap to image coordinates, our formulation operates directly in 3D

Table 1. **Reconstruction performance.** Our method consistently improves both all-age and adult-only initializations across all metrics by a large margin, with adult-only models becoming competitive with all-age methods. Δ: improvement over initialization.

(a) **In-the-wild reconstruction on Relative Human test.** $*$: trained on the train set.

Method	**2D** (↑) $mPCKh^{0.6}$	**PCRD**$^{0.2}$ (↑) overall	adult	teen	kid	baby	**Age** (↑) F1	**Gender** (↑) F1
*BEV** [52]	*74.78*	*69.19*	*70.65*	*68.27*	*65.35*	*61.96*	*30.32*	*0.00*
Multi-HMR [7]	65.39	59.79	60.73	64.58	56.45	46.3	23.29	34.83
+ **Ours**	78.84	66.11	66.92	65.59	**67.08**	**57.21**	48.57	81.11
Δ	+13.45	+6.32	+6.19	+1.01	+10.63	+10.91	+25.28	+46.28
CameraHMR [38]	64.69	59.59	59.80	67.49	46.59	30.71	0.00	0.00
+ **Ours**	**81.06**	**67.24**	**67.74**	**70.36**	65.60	51.09	**48.75**	**82.13**
Δ	+16.37	+7.65	+7.94	+2.87	+19.01	+20.38	+48.75	+82.13

(b) **3D reconstruction on CMU panoptic-toddler.**

Method	**Root** MPJPE (↓ mm)	PCK (↑ %)	**Joint-PA** MPJPE (↓ mm)	PCK (↑ %)
AiOS [50]	162.39	54.15	723.14	7.78
SAT-HMR [49]	153.88	56.40	654.87	5.50
Multi-HMR [7]	102.15	84.50	263.78	41.89
+ **Ours**	**92.52**	**85.50**	**223.13**	**45.12**
Δ	-9.63	+1.00	-40.66	+3.23
CameraHMR [38]	149.52	60.18	658.90	5.53
+ **Ours**	119.93	74.41	348.03	23.32
Δ	-29.60	+14.23	-310.86	+17.79

space, enforcing relational consistency between subjects and avoiding independent local minima. The optimization proceeds in stages to prevent degenerate solutions: we first optimize only translation τ to resolve coarse depth placement, then optimize $\{\tau, \phi, \beta\}$ to refine global orientation and shape attributes while preserving stable positioning, and finally optimize all parameters $\{\tau, \phi, \beta, \theta\}$ to recover detailed pose.

Scene cues To place all individuals within a coherent scene, we incorporate spatial relationships by extending the depth ordering loss from [52], denoted L_{depth}, to continuous pseudo ground-truth depth values. This loss encourages people predicted to lie on the same depth plane to be close together, while separating those on different planes. While [52] relies on manually annotated depth levels, we instead make use of an off-the-shelf metric depth map estimator [42], which we found to provide consistent ordinal depth cues. We estimate the median depth of each person using the corresponding segmentation mask M predicted using [43], which we find more robust than relying on J due to potential occlusions and errors in D. In addition to depth ordering, we further stabilize the optimization by regularizing with the previous estimates Θ_{t-1}, ensuring smooth refinement from the initialization Θ_{init}.

In sum, the weighted loss function for our Anny-Fit optimization is:

$$\mathcal{L} = \lambda_{2D}\mathcal{L}_{2D} + \lambda_{dense}\mathcal{L}_{dense} + \lambda_{shape}\mathcal{L}_{shape} + \lambda_{init}\mathcal{L}_{init} + \lambda_{depth}\mathcal{L}_{depth}$$

where the λ coefficients balance the contributions of the individual losses, which are also flexibly adjusted across optimization stages (see Supp. Mat for more details).

4. Experiments

Initialization. To validate our optimization-based framework, we initialize it with two complementary feedforward methods: Multi-HMR [2, 7] and CameraHMR [38]. Multi-HMR uses the Anny model and is trained only on synthetic data (Anny-One [7]), whereas CameraHMR is a detection-based, single-person, adult-only model trained on a mixture of synthetic data and real-world pseudo-ground-truth fits. Multi-HMR [7] outputs Anny body parameters directly, while for CameraHMR we fit Anny parameters to the SMPL meshes [31]. For both initial predictions we use bounding boxes, for the bottom-up method Multi-HMR we force predictions using the closest patch prediction of the estimated nose keypoint.

Evaluation benchmarks and metrics. We compare to prior work on the in-the-wild all-age dataset Relative Human [52] and evaluate 3D recovery on the 5 toddler sequences of the CMU Panoptic dataset [21] as well as in Hi4D [61]. For 2D evaluation, we use mean Percentage of Correct Keypoints ($mPCK^{0.6}$), Percentage of Correct Depth Relations ($PCDR^{0.2}$), F1 for bounding-box detection, and age and gender prediction. For 3D evaluation we measure MPJPE on matched joints and PCK @$15cm$ to account for unmatched joints. We consider both the per-person root aligned and Procrustes alignment (i.e. Joint-PA) for all people as one to account for inter-person accuracy [37]. We include more details in the Supp. Mat.

Comparison to state-of-the-art Table 1a reports results on the Relative Human test and comparison with existing methods. Anny-Fit provides substantial improvements over zero-shot initializations across all metrics, in several cases matching or surpassing BEV, the current SOTA, despite BEV being trained directly on the dataset. The results also highlight the coupled nature of the task: improvements in shape attribute estimation (age and gender F1) directly reduce depth–shape ambiguity, enabling more reliable multi-person depth ordering. This stronger spatial grounding, in turn, supports more accurate pose recovery as reflected by 2D reprojection gains. These findings are further supported by the 3D evaluation on CMU-Toddler reported in Table 1b, which measures reconstruction accuracy on real-world multi-age sequences. Anny-Fit improves single-person pose recovery (root-aligned metrics) and multi-person relative placement (joint-PA metrics).

Overall, we demonstrate that Anny-Fit consistently benefits different model types. For the specialized all-age model (Multi-HMR), improvements stem from better alignment between subject appearance and 3D shape, mitigating limitations of its synthetic training distribution. At the same time, Anny-Fit shows that predictions from a

Table 2. **Method ablation study.** Anny-Fit main component analysis on the Relative Human validation 'has child' subset. indicates the default setting. O: Multi-person optimization, S: VLM shape, D: depth, RD: root depth.

Method	2D	PCRD$^{0.2}$					Age	Gender
	$mPCKh^{0.6}$	overall	adult	teen	kid	baby	F1	F1
Multi-HMR [7]	60.99	62.66	61.22	77.10	62.85	56.07	28.55	41.95
+ O	71.76	60.09	59.47	69.73	59.33	56.82	36.05	77.41
+ O + S	76.28	59.95	61.37	68.83	58.16	58.28	59.70	84.53
+ O + D	79.21	63.37	63.21	72.45	63.94	54.17	30.61	38.38
+ O + S + RD	78.22	64.86	66.55	71.52	**63.59**	**60.79**	**57.45**	**84.91**
+ O + S + D	**79.22**	**65.13**	**66.67**	**72.99**	63.31	58.86	56.78	83.75
CameraHMR [38]	43.84	50.82	46.41	66.12	52.87	34.06	18.49	38.04
+ O	79.09	55.01	50.92	71.33	55.88	38.22	22.47	39.16
+ O + S	80.90	55.78	53.69	64.85	57.82	42.45	**57.45**	**83.45**
+ O + D	79.86	59.43	57.30	68.68	61.38	45.91	43.94	41.27
+ O + S + RD	**82.27**	66.15	62.43	77.23	68.45	**58.47**	56.99	82.09
+ O + S + D	81.62	**67.55**	**66.06**	**78.87**	**70.65**	53.77	56.52	81.43

Table 3. **Shape attribute performance.** Results of varying VLM models on the Relative Human validation 'has child' subset. indicates the default setting.

Method	Age F1					Gender
	overall	adult	teen	kid	baby	F1
SmolVLM-Instruct [33]	43.06	28.05	34.46	52.31	57.41	90.23
ViP Llava 13B [8]	46.64	62.44	**62.98**	19.23	41.89	83.87
Llama3 Llava-NeXt 8B [26]	52.46	70.18	28.85	60.58	50.25	78.80
Qwen2.5 VL 3B [1]	63.57	81.58	36.49	66.34	**69.86**	**92.26**
Qwen2.5 VL 7B [1]	**67.23**	**85.98**	51.57	**70.29**	61.08	92.03

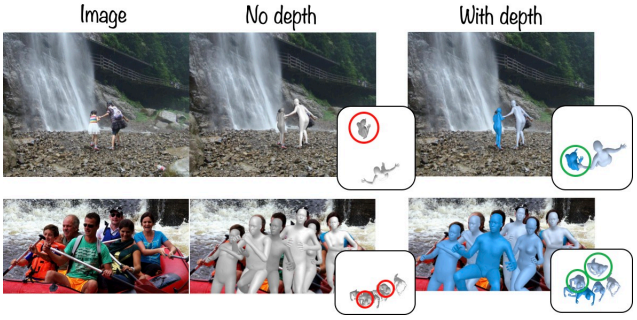

Figure 5. **Effect of depth loss.** Adding a depth-based loss from an expert model preserves the relative depth relationships between people. All results use the same initialization and shape prediction. Circles denote incorrect (red) and correct (green) placement.

non-specialized, adult-only model (CameraHMR) can be adapted to the all-age setting in a training-free manner, achieving competitive performance with specialized methods.

Figure 4 further visualizes these capabilities on challenging in-the-wild scenes featuring high shape variance, extreme poses, and close interactions. Anny-Fit leverages and generalizes SOTA capabilities; it refines relative depth ordering in both multi-age arrangements (rows 1, 4) and adult-only scenes (row 3). By building upon accurate pose estimators such as CameraHMR (rows 2, 3), Anny-Fit captures extreme poses, and using a stronger detectors can increase recall in certain scenes with challenging poses or heavy occlusions (rows 1, 2). Finally, by using the more representative Anny body model, Anny-Fit avoids shape artifacts common in SMPL-A methods such as BEV, where children may appear as oversized infants (row 2). These results highlights the flexibility of Anny-Fit, which directly benefits from advances in HMR methods even when they are not designed for all-age reconstruction.

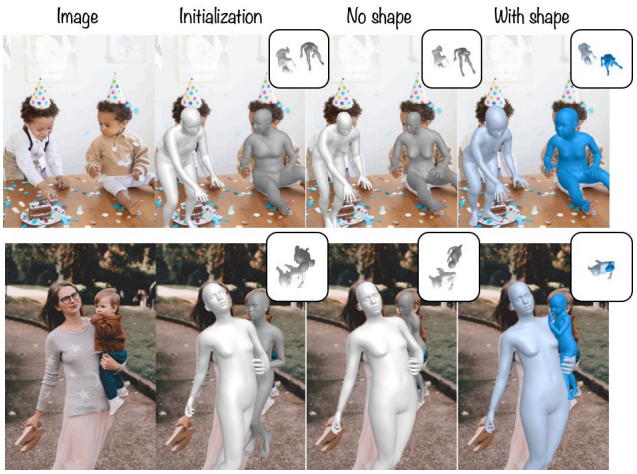

Figure 6. **Effect of shape.** Top: Incorrect shape initialization prevents the optimization from converging. Bottom: Accurate shape initialization resolves depth ordering.

4.1. Ablation experiments

We study each component of our method on the Relative Human validation set. To counter the over-representation of adults, we report on a 'has child' subset containing all images with at least one non-adult (see Supp. Mat.).

Component ablation study. Table 2 summarizes the contributions of individual components. Adding weak 2D keypoint supervision improves $mPCK^{0.6}$ and yields small gains in shape, but has limited effect on depth. In contrast, depth supervision provides a large boost to $PCDR^{0.2}$, especially for CameraHMR initializations where non-adult predictions are severely miss-scaled. The depth loss is crucial for bringing these cases back into a valid scale (see Fig. 5).

Integrating VLM-estimated shape attributes — both at initialization and as a shape loss — has the largest impact, producing substantial gains in age and gender classification (+30 and +40 F1, respectively). It also improves depth and pose accuracy. As shown in Fig. 6, when the initial shape is far from the true one, optimization cannot recover it without shape guidance. Consistent with prior work [52, 54] on the depth-scale ambiguity, using the correct shape leads to more reliable relative depth. We further evaluate two depth losses: a learned affine transformation of the per-person predicted median depth (RD) and the final relative depth ordering loss (D). Both improve multi-person depth consistency, with the ordering loss performing best overall. The full model, combining all components, achieves the strongest overall performance across metrics, underscoring

Table 4. **Training feedforward models with pseudo ground-truth**. Comparing Multi-HMR retrained with different data.

(a) **Relative Human test.**

Data	2D $mPCKh^{0.6}$	PCRD$^{0.2}$ overall	adult	teen	kid	baby	Age F1	Gender F1
Anny-One [6]	62.70	63.42	63.93	71.40	59.70	44.17	24.47	33.60
Anny-One + [38] fits	59.75	52.09	53.10	59.58	43.17	25.93	11.75	29.55
Anny-One + **our fits**	**70.18**	**68.68**	**69.03**	**76.83**	**66.42**	**57.52**	**42.96**	**81.32**
Δ	+7.48	+5.26	+5.10	+5.43	+6.72	+13.35	+18.49	+47.72

(b) **Hi4D test.**

Data	MPJPE (↓ mm)	Joint-PA MPJPE (↓ mm)
Anny-One [6]	91.5	86.7
Anny-One + [38] fits	80.9	81.8
Anny-One + **our fits**	**80.1**	**79.5**
Δ	-11.4	-7.2

the necessity of jointly modeling shape, depth, and pose in the all-age setting.

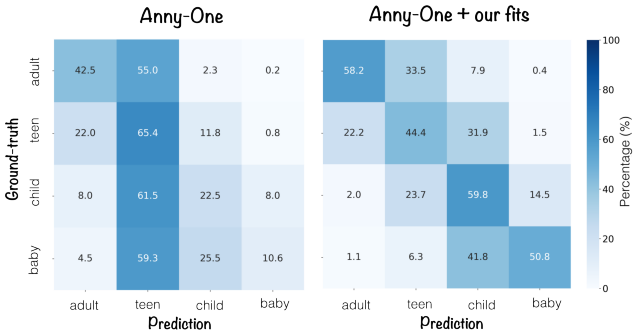

Figure 7. **Age confusion matrix on Relative Human test.** After retraining with our fits, age alignment improves.

VLM shape attribute estimation. Table 3 reports zero-shot shape-attribute predictions from open-source VLMs that fit on a single H100 GPU, using prompt aligned with Relative Human's class definitions. Gender is predicted accurately across all models, suggesting that this attribute is well captured. In contrast, age estimation shows large variability: performance is strongest for adults, while most errors occur between neighboring age ranges, indicating that finer age distinctions remain challenging.

Pseudo-ground truth generation. Finally, we assess whether our high-quality fits can be used to train feedforward HMR methods. To this end, we processed 30k images from the MS-COCO [28] training set, which provides diverse, in-the-wild scenes spanning a broad range of ages. We trained Multi-HMR using various combinations of synthetic data, our optimized fits, and fits generated by the optimization method CamSimplify [38]. All models are trained for 600K steps at input resolution of 672×672.

We report results in Table 4. On Relative Human, adding our fits mitigates the limitations of synthetic-only training and yields large improvements in shape estimation (+18, +47), 2D reprojection (+7), and depth ordering (+5–13). Fig. 7 illustrates how our fits improve alignment of predicted age after training. In contrast, adding CamSimplify fits degrades performance, showing that gains stem from the quality of the pseudo-GT rather than simply more data. Results on Hi4D confirm the same trend: our fits provide larger improvements than CamSimplify even on an adult-only dataset. Overall, these findings highlight the importance of high-quality, camera-consistent multi-person fits and demonstrate the potential of Anny-Fit to enhance a base model when full supervision is limited.

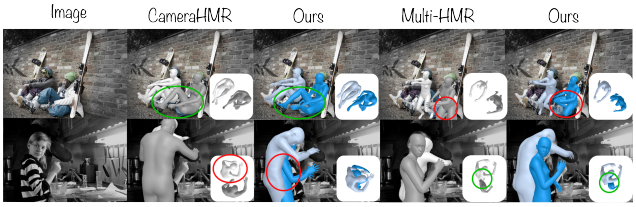

Figure 8. **Example of failure cases**. Top: pose of legs, legs usually have low keypoint confidence. Bottom: interpenetration because of initial position and not enough guiding keypoints.

5. Limitations

As an optimization-based method, our approach is highly dependent on the quality of both the initial mesh parameters and expert predictions. Fig. 8 shows failure cases. For example, errors in the experts –such as low-confidence keypoints or misclassified shape attributes–can prevent pose and shape convergence. Similarly, poor global-position initialization may cause optimization to stall or produce interpenetration when cues are weak. Future work could tackle this challenge, as explored previously by [34, 35, 37]. These issues highlight remaining challenges in complex multi-person scenes.

6. Conclusion

We introduced AnyFits, a robust multi-person optimization framework that jointly fits all individuals in the camera coordinate system using complementary expert signals, including VLM-derived semantic attributes. AnyFits improves spatial consistency, pose accuracy, and shape estimation across challenging benchmarks in all-age scenarios where standard per-person and traditional adult-only methods fail. We further demonstrated that AnyFits can generate high-quality pseudo-ground-truth annotations at scale, enabling the training of feedforward HMR models that achieve strong performance while predicting semantically meaningful shape parameters. Together, these results highlight the potential of expert-guided optimization to bridge the gap between adult-only models and real-world, all-age human reconstruction.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 4, 7

[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, 3, 6

[3] Siyuan Bian, Chenghao Xu, Yuliang Xiu, Artur Grigorev, Zhen Liu, Cewu Lu, Michael J Black, and Yao Feng. Chatgarment: Garment estimation, generation and editing via large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2924–2934, 2025. 4

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 2

[5] Laura Bravo-Sánchez, Jaewoo Heo, Zhenzhen Weng, and Kuan-Chieh Wang. Ask, pose, unite: Scaling data acquisition for close interaction meshes with vision language models. In *Synthetic Data for Computer Vision Workshop@ CVPR 2025*, 2025. 2

[6] Romain Brégier, Fabien Baradel, Thomas Lucas, Salma Galaaoui, Matthieu Armando, Philippe Weinzaepfel, and Grégory Rogez. Condimen: Conditional multi-person mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3880–3890, 2025. 2, 3, 8

[7] Romain Brégier, Guénolé Fiche, Laura Bravo-Sánchez, Thomas Lucas, Matthieu Armando, Philippe Weinzaepfel, Grégory Rogez, and Fabien Baradel. Human mesh modeling for anny body. *arXiv preprint arXiv:2511.03589*, 2025. 2, 3, 4, 6, 7, 1

[8] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 7

[9] Vasileios Choutas, Lea Müller, Chun-Hao P Huang, Siyu Tang, Dimitrios Tzionas, and Michael J Black. Accurate 3d body shape regression using metric and semantic attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2718–2728, 2022. 2, 3

[10] Andrey Davydov, Anastasia Remizova, Victor Constantin, Sina Honari, Mathieu Salzmann, and Pascal Fua. Adversarial parametric pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10997–11005, 2022. 3

[11] Delmas, Ginger and Weinzaepfel, Philippe and Lucas, Thomas and Moreno-Noguer, Francesc and Rogez, Grégory. PoseScript: Linking 3D Human Poses and Natural Language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[12] Delmas, Ginger and Weinzaepfel, Philippe and Moreno-Noguer, Francesc and Rogez, Grégory. PoseFix: Correcting 3D Human Poses with Natural Language. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2023.

[13] Delmas, Ginger and Weinzaepfel, Philippe and Moreno-Noguer, Francesc and Rogez, Grégory. PoseEmbroider: Towards a 3D, Visual, Semantic-aware Human Pose Representation. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[14] Yao Du, Qiang Zhai, Weihang Dai, and Xiaomeng Li. Teach clip to develop a number sense for ordinal regression. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 5

[15] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2093–2103, 2024. 3, 4

[16] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 3

[17] 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. 2

[18] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Raphael Weinberger, and A Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 3

[19] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. Learning an infant body model from rgb-d data for accurate full body motion analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 792–800. Springer, 2018. 3

[20] Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1011–1021, 2024. 3

[21] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 6, 3

[22] Yiming Shi Chenyi Guo Ji Wu Kaiwen Wang, Kaili Zheng. Towards metric-aware multi-person mesh recovery by jointly optimizing human crowd in camera space. *arXiv preprint arXiv:2511.13282*, 2025. 3

[23] 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. 3

[24] 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. 3

[25] 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, 3

[26] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024. 7

[27] 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 in Computer Vision*, 2022. 3

[28] 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. 8

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3, 6

[32] Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, and Ziwei Liu. Dposer-x: Diffusion model as robust 3d whole-body human pose prior. *arXiv preprint arXiv:2508.00599*, 2025. 3

[33] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 7

[34] Marko Mihajlovic, Siwei Zhang, Gen Li, Kaifeng Zhao, Lea Muller, and Siyu Tang. Volumetricsmpl: A neural volumetric body model for efficient interactions, contacts, and collisions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5060–5070, 2025. 8

[35] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9990–9999, 2021. 2, 8

[36] Lea Muller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9687–9697, 2024. 2

[37] Lea Muller, Vickie Ye, Georgios Pavlakos, Michael J. Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3D social interaction from images. 2024. 3, 6, 8

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

[39] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 2, 3

[40] 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. 3

[41] Pexels. Pexels stock photos. https://www.pexels.com/, 2025. 3

[42] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. 6

[43] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 6

[44] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 10–15, 2015. 3

[45] 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. 1

[46] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, 2020. 3

[47] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L Yuille. Deep regression forests for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2304–2313, 2018. 3

[48] Stephan Streuber, M Alejandra Quiros-Ramirez, Matthew Q Hill, Carina A Hahn, Silvia Zuffi, Alice O'Toole, and Michael J Black. Body talk: Crowdshaping realistic 3d avatars with words. *ACM Transactions on Graphics (TOG)*, 35(4):1–14, 2016. 3

[49] Chi Su, Xiaoxuan Ma, Jiajun Su, and Yizhou Wang. Sathmr: Real-time multi-person 3d mesh estimation via scale-adaptive tokens. In *Proceedings of the IEEE/CVF Confer-*

ence on Computer Vision and Pattern Recognition, 2025. 3, 6

[50] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, et al. Aios: All-in-one-stage expressive human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 6

[51] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, One-stage, Regression of Multiple 3D People. In *ICCV*, 2021.

[52] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 2, 3, 6, 7, 1

[53] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8001–8013, 2023. 2

[54] Nicolas Ugrinovic, Adria Ruiz, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Body size and depth disambiguation in multi-person reconstruction from single images. In *2021 International Conference on 3D Vision (3DV)*, pages 53–63. IEEE, 2021. 2, 3, 7

[55] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14644–14654, 2023. 2

[56] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1148–1159, 2025. 2, 3

[57] Zhenzhen Weng, Laura Bravo-Sánchez, Zeyu Wang, Christopher Howard, Maria Xenochristou, Nicole Meister, Angjoo Kanazawa, Arnold Milstein, Elika Bergelson, Kathryn L. Humphreys, Lee M. Sanders, and Serena Yeung-Levy. Artificial intelligence–powered 3d analysis of video-based caregiver-child interactions. *Science Advances*, 11(8): eadp4422, 2025. 3

[58] Yan Xia, Xiaowei Zhou, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Reconstructing humans with a biomechanically accurate skeleton. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5355–5365, 2025. 2

[59] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 3

[60] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose+: Vision transformer foundation model for generic body pose estimation. *arXiv preprint arXiv:2212.04246*, 2022. 5

[61] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 6

[62] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2148–2157, 2018. 3

[63] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023. 2

[64] He Zhang, Chentao Song, Hongwen Zhang, and Tao Yu. Metrichmr: Metric human mesh recovery from monocular images. *arXiv preprint arXiv:2506.09919*, 2025. 2, 3

[65] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 5

[66] Rui Zhu, Xingyi Yang, Yannick Hold-Geoffroy, Federico Perazzi, Jonathan Eisenmann, Kalyan Sunkavalli, and Manmohan Chandraker. Single view metrology in the wild. In *European Conference on Computer Vision*, pages 316–333. Springer, 2020. 2

[67] Nikolaos Zioulis and James F O'Brien. Kbody: Towards general, robust, and aligned monocular whole-body estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6215–6225, 2023. 2