Towards Metric-Aware Multi-Person Mesh Recovery by Jointly Optimizing Human Crowd in Camera Space

Kaiwen Wang[1*] Kaili Zheng[1*] Yiming Shi[1] Chenyi Guo[1✉] Ji Wu[1,2,3✉]
[1]Department of Electronic Engineering, Tsinghua University
[2]College of AI, Tsinghua University
[3]Beijing National Research Center for Information Science and Technology
[*] Equal contribution [✉] Corresponding author
{wkw23, zkl25, shiym23}@mails.tsinghua.edu.cn
{guochy, wuji_ee}@mail.tsinghua.edu.cn

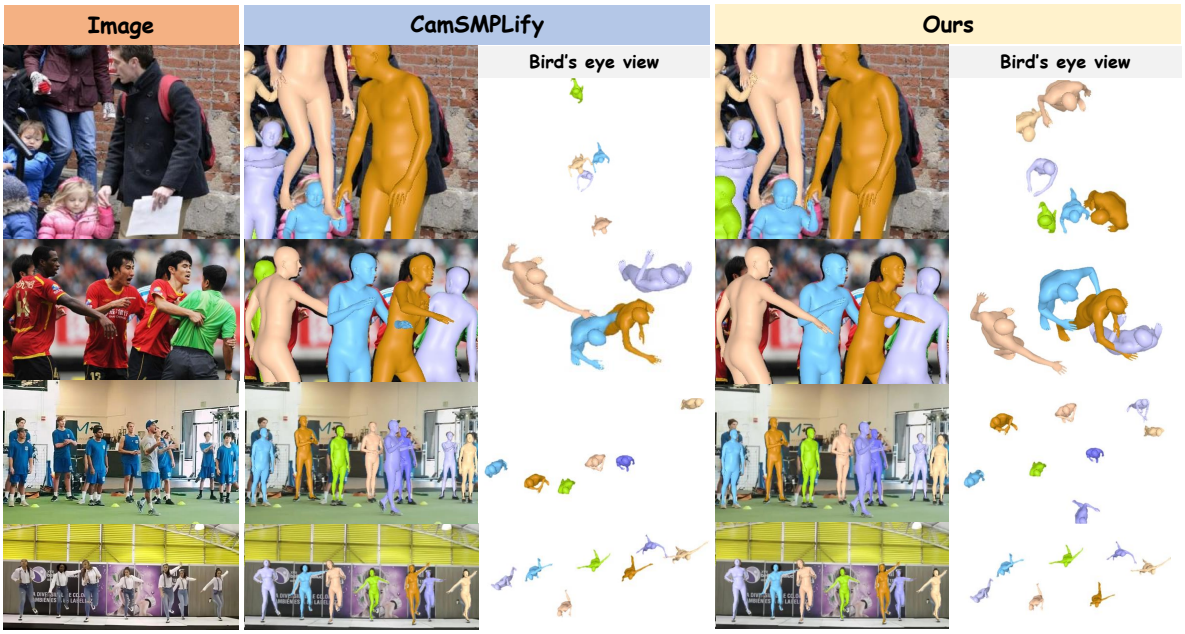

Figure 1. CamSMPLify [41] fits each person independently, causing height and spatial inconsistencies. Our DTO jointly optimizes human translations from height priors and depth cues, ensuring a coherent scene reconstruction.

Abstract

*Multi-person human mesh recovery from a single image is a challenging task, hindered by the scarcity of in-the-wild training data. Prevailing in-the-wild human mesh pseudo-ground-truth (pGT) generation pipelines are single-person-centric, where each human is processed individually without joint optimization. This oversight leads to a lack of scene-level consistency, producing individuals with conflicting depths and scales within the same image. To address this, we introduce **D**epth-conditioned **T**ranslation **O**ptimization (DTO), a novel optimization-based method that jointly refines the camera-space translations of all individuals in a crowd. By leveraging anthropometric priors on human height and depth cues from a monocular depth estimator, DTO solves for a scene-consistent placement of all subjects within a principled Maximum a posteriori (MAP) framework. Applying DTO to the 4D-Humans dataset, we construct **DTO-Humans**, a new large-scale pGT dataset of 0.56M high-quality, scene-consistent multi-person images, featuring dense crowds with an average of 4.8 persons per image. Furthermore, we propose **Metric-Aware HMR**, an*

end-to-end network that directly estimates human mesh and camera parameters in metric scale. This is enabled by a camera branch and a relative metric loss that enforces plausible relative scales. Extensive experiments demonstrate that our method achieves state-of-the-art performance on relative depth reasoning and human mesh recovery. Code is available at: https://github.com/gouba2333/MA-HMR.

1. Introduction

Recovering 3D human pose and shape from a single image is a fundamental challenge in computer vision with profound implications for applications such as autonomous driving [51], augmented and virtual reality [56], sports analysis [33]. While much progress has focused on individuals, real-world scenes are typically composed of multiple people. Challenges like heavy occlusions and depth ambiguities prevalent in multi-person scenes make holistic scene understanding a hard problem to solve.

While the construction of in-the-wild 3D human datasets is crucial for training deep learning models, acquiring such data at a large scale remains exceptionally difficult. To overcome this, the community has widely adopted a paradigm of pseudo-ground-truth (pGT) generation [20, 24, 38, 41]. However, current pipelines are single-person-centric. They optimize each individual in isolation and simply assemble the results. This lack of joint optimization ignores scene-level spatial relationships, yielding physically implausible layouts with incorrect depth ordering and inconsistent relative scales (see Fig. 1, mid). These inconsistent pGTs limit the potential of powerful end-to-end regression frameworks [2, 44–46] which are capable of learning complex human-human and human-environment relations, and thus place a data bottleneck in the field of multi-person mesh recovery.

To resolve scene-wide inconsistencies and put people in their place, we model the problem by mirroring human-like inference, which relies on two key information sources: prior knowledge of human height and visual cues of relative depth. We leverage off-the-shelf models [27] to predict age and gender, from which we derive a statistically-informed height distribution for each individual [6, 17]. For the visual cues, we utilize a monocular depth estimator [59] to extract rich geometric information. We then propose a unified framework to integrate these information, named **D**epth-conditioned **T**ranslation **O**ptimization (DTO). DTO solves a joint Maximum a posteriori (MAP) problem, where the height distribution acts as the prior. The likelihood term is conditioned on two robust depth constraints: inter-human depth relations for positioning and intra-human depth scale for structural consistency. The resulting optimization finds the globally optimal human translations that make the entire crowd spatially consistent in the scene (see Fig. 1, right).

We then address the data bottleneck by applying DTO to 2M images from the 4D-Humans dataset [1, 21, 32, 54] and leverage the posterior probability derived from our DTO framework to filter out poorly optimized or inconsistent scenes. This yields **DTO-Humans**, a large-scale pGT dataset comprising 0.56M high-quality, scene-consistent multi-person images. With 2.7M person instances in dense scenes (avg. 4.8 persons/image), this dataset provides consistent, physically-grounded supervision that has long been a missing piece in training robust multi-person models.

Building upon this dataset, we further propose Metric-Aware HMR, an end-to-end network that directly regresses human mesh parameters in metric scale. A critical challenge in metric recovery is the unknown camera intrinsic parameters, which are intricately coupled with the perceived scale of individuals and the overall scene composition. An end-to-end architecture is uniquely suited to resolve this ambiguity, as it can process global image information. Our network materializes this insight by incorporating a camera branch with the HMR model and simultaneously predicts the camera's field of view (FoV) and human mesh parameters. To guide this joint prediction, we introduce a novel relative metric loss, which explicitly penalizes implausible real-world size relationships among individuals. By learning to predict all parameters concurrently from the full image context, our model learns the complex interplay between scene geometry and individual human scales.

In summary, our contributions are as follows:

- We propose DTO, a novel MAP optimization framework that jointly optimizes multi-person camera-space translations using both inter-human and intra-human depth constraints to ensure scene-level consistency.
- We construct DTO-Humans, a large-scale pGT dataset with high scene consistency, paving the way for training more robust multi-person models.
- We design an end-to-end network, Metric-Aware HMR, featuring a dedicated camera branch and a relative metric loss to achieve true metric-scale mesh recovery.
- Extensive experiments show that our method achieves state-of-the-art performance on multi-person 3D reconstruction benchmarks, including RH [48], 3DPW [50], CMU-Panoptic [19], Hi4D [60], and MuPoTS [37].

2. Related work

2.1. Multi-person Human Mesh Recovery

Multi-person human mesh recovery aims to regress all human meshes from a single RGB image. Existing methods can be categorized into multi-stage and one-stage approaches. Multi-stage approaches [4, 11, 30, 41] detect all humans in the image and apply a single-person HMR model to each of them separately. In contrast, single-stage paradigms process the entire image at once to jointly recover all individuals, which allow the model to capture

human-human and human-environment interactions. Early methods like ROMP [47] and BMP [61] rely on person-center heatmaps. More recently, DETR [5, 28, 34]-based frameworks [2, 44–46] have shown strong performance by using human queries to decode mesh parameters from feature maps. While these end-to-end models have the potential for holistic scene understanding, they highly depend on the quantity and quality of training data.

2.2. Multi-person pGT Generation

The scarcity of in-the-wild 3D human data has made pseudo-ground-truth (pGT) generation a cornerstone of modern HMR [20, 24, 25, 38, 41]. This paradigm, uses optimization-based methods like SMPLify [35] to create 3D labels for large 2D datasets [1, 32], which then supervise regression networks. However, existing pipelines optimize each person in isolation, leading to scene-inconsistent pGT datasets that directly limit model performance.

To improve pGT quality in multi-person scenes, various constraints have been explored. Some methods introduce constraints like a physics-based inter-penetration loss to prevent mesh intersections or a depth-ordering loss to penalize incorrect occlusions [18, 22]. While effective for scenes where people are physically close or overlapping, these approaches have limited utility in common scenarios where people are spatially separated. Recent advances in monocular depth estimation, particularly powerful models like Depth Anything [58, 59], provide rich geometric cues for scene understanding. While prior work has used depth features to estimate single-person translation [52], our work is the first to leverage these relative depth cues within a joint optimization framework to enforce consistency among multiple people, enabling the creation of scene-consistent pGT.

Anthropometric priors offer another constraint. Prevailing works ensure realistic body shapes by regularizing shape parameters [24, 35, 43] or associating shape with text prompts [8, 53]. However, these strategies are limited by their datasets, which often lack diversity in body size and hard to generalize to the full range of real-world human heights. Our approach leverage the progress in age and gender prediction [26, 27] and establish a statistically-informed [6, 17] height prior for each individual.

3. Method

3.1. Preliminaries

Human Model We adopt the unified body model from AGORA [42], which supports varying ages by blending the SMPL [35] and SMIL [12] templates. It is parameterized by a pose vector $\theta \in \mathbb{R}^{24\times3}$ and a shape vector $\beta \in \mathbb{R}^{11}$, where the final component of β controls the age-related blending. The model function $\mathcal{M}(\theta, \beta)$ outputs a 3D mesh $V \in \mathbb{R}^{6890\times3}$. To place the human within the camera coordinate system, we apply a 3D translation vector $T = (x, y, z)$, resulting in the final 3D points $P = V + T$.

Camera Model We use a perspective camera model to project a 3D point $P = (X, Y, Z)$ from the camera coordinate system onto a 2D point p on the image plane. We make standard simplifying assumptions that the camera has no radial distortion, and the principal point (c_x, c_y) is at the image center. Following prior work [41], instead of directly estimating the camera's focal length f, we predict its vertical FoV v. The focal length can then be derived using the image height H: $f = \frac{H}{2\cdot\tan(v/2)}$. The full projection function is then: $p = f \cdot \frac{(X,Y)}{Z} + (c_x, c_y)$.

Monocular Depth Estimation We leverage an off-the-shelf monocular depth estimator [59] to provide geometric cues. This model outputs a relative depth map D_{rel}, which is related to the true metric depth D_{metric} by an unknown scale s_d and shift t_d: $D_{\text{metric}} = s_d \cdot D_{\text{rel}} + t_d$. Our DTO framework is designed to solve for these global affine parameters (s_d, t_d) for the entire scene, thereby grounding all individuals in a consistent metric space.

3.2. DTO Framework

To address the scene-inconsistency of independent per-person estimations, we introduce Depth-conditioned Translation Optimization (DTO). DTO formulates the task of refining all individuals' camera-space placements as a joint optimization problem. As illustrated in Fig. 2, our approach leverages geometric cues from monocular depth and statistical human priors within an MAP problem to find a globally optimal solution. By applying DTO to a large collection of in-the-wild images, we create DTO-Humans, a large-scale dataset for scene-consistent human reconstruction.

3.2.1. Initial Per-Person Estimation

We obtain initial per-person estimates via a segmentation-enhanced process [23], using Detectron2 [55] to generate person bounding boxes and segmentation masks. For each masked person, we employ CameraHMR [41] to predict their initial mesh. Subsequently, we use 2D keypoints detected by ViTPose [57] to filter out inaccurate mesh estimations through a matching process (details in the supplementary material). We utilize HumanFoV [41] to estimate the camera instrinsics to place human meshes in camera space.

3.2.2. DTO Formulation and Core Components

Our DTO framework ensures scene consistency by jointly optimizing for a global affine depth transformation, parameterized by a scale s_d and a shift t_d. We formulate this as an MAP problem, which requires a robust prior and a way to connect it to the variables (s_d, t_d) using geometric evidence. This section details the core components of our formulation.

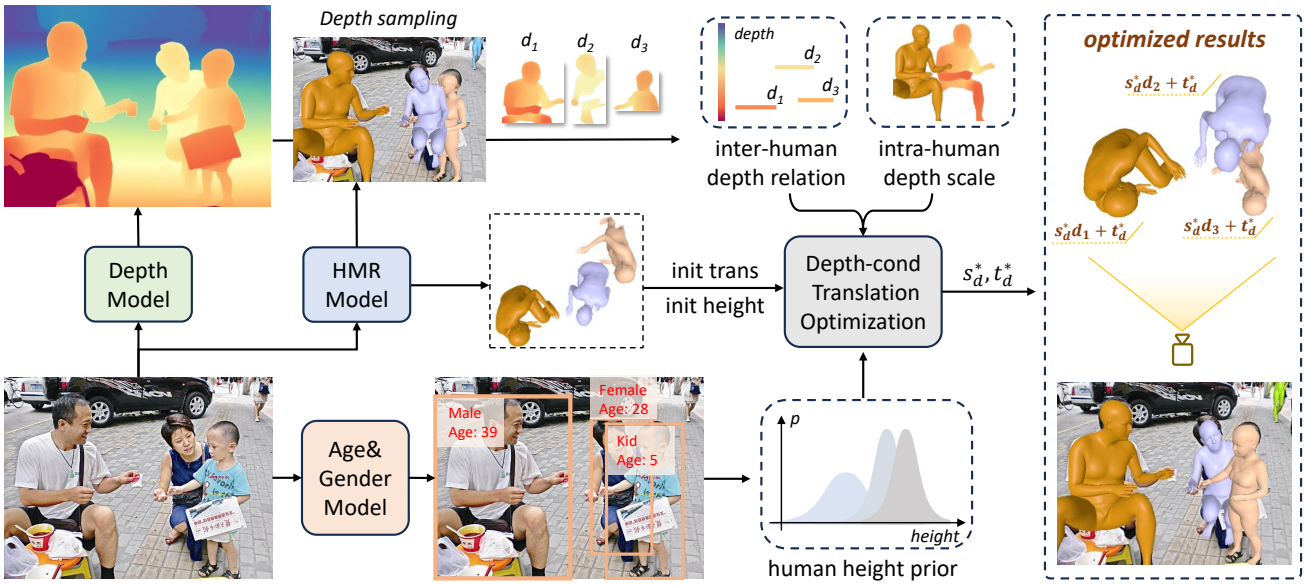

Figure 2. Overview of the DTO Framework. An input image is processed through three parallel streams: an HMR Model provides initial meshes; Age & Gender Model informs a statistical height prior; and Depth Model generates a relative depth map. From the initial meshes and the depth map, we extract the inter-human depth relation and intra-human depth scale. DTO integrates these components into an MAP problem to solve for a global affine transformation, outputting a scene-consistent arrangement of all individuals.

Height Prior We establish an anthropometric prior on each person's height. We utilize the expert model [27] to estimate the age and gender of each individual and define the height prior for the person i as a Gaussian distribution, $P_i(h_i) = \mathcal{N}(h_i|\mu_i, \sigma_i)$, where the mean μ_i and standard deviation σ_i are derived from demographic statistics [6, 17]. Details are available in the supplementary material.

Inter-human Depth Relation To enforce consistent relative positioning, we extract geometric cues from a relative depth map D_{rel} generated by the depth model [59]. For each person i, we define their representative depth value, d_i, as the mean depth sampled from D_{rel} at the 2D projected locations of all their visible mesh vertices.

Intra-human Depth Scale A global affine transformation ($D_{\text{metric}} = s_d D_{\text{rel}} + t_d$) without proper constraints can lead to physically implausible solutions. To prevent this, we introduce the intra-human depth scale to define a plausible range for the global scale factor s_d. For each person, we compute the linear regression slope X_i and the correlation coefficient w_i between their internal mesh depth and the values from the relative depth map. We then use the correlation coefficients as weights and compute a weighted average of these individual slopes as a scale estimate:$X = \frac{\sum_{i=1}^{K} w_i X_i}{\sum_{i=1}^{K} w_i}$. From this estimate, we establish a plausible range for s_d by defining its lower and upper bounds, $X_{\min}$ and $X_{\max}$, as $X_{\min} = \alpha_1 X$ and $X_{\max} = \alpha_2 X$. The selection of hyperparameters α_1, α_2 is detailed in the supplementary material.

3.2.3. DTO Objective and Solution

With these components, we tend to find the optimal global affine transformation parameters (s_d, t_d) that make the corrected height of every person most probable under their respective priors. The corrected height h_i is a transformation of the initial T-pose height $\hat{h}_i$ and predicted depth $\hat{z}_i$:

$$h_i = \hat{h}_i \cdot \frac{s_d d_i + t_d}{\hat{z}_i} \tag{1}$$

This leads to the following MAP optimization problem:

$$\max_{s_d, t_d} \prod_{i=1}^{K} \mathcal{N}\left(\hat{h}_i \cdot \frac{s_d d_i + t_d}{\hat{z}_i} \middle| \mu_i, \sigma_i\right) \tag{2}$$

Taking the negative log-likelihood and incorporating our bounded intra-human scale constraint converts this into a constrained minimization problem:

$$\min_{s_d, t_d} \sum_{i=1}^{K} \frac{\left(\hat{h}_i \cdot \frac{s_d d_i + t_d}{\hat{z}_i} - \mu_i\right)^2}{\sigma_i^2} \quad \text{s.t. } X_{\min} \le s_d \le X_{\max} \tag{3}$$

This objective is a convex quadratic program with analytical solution via the Karush-Kuhn-Tucker conditions. The full derivation is provided in the supplementary material.

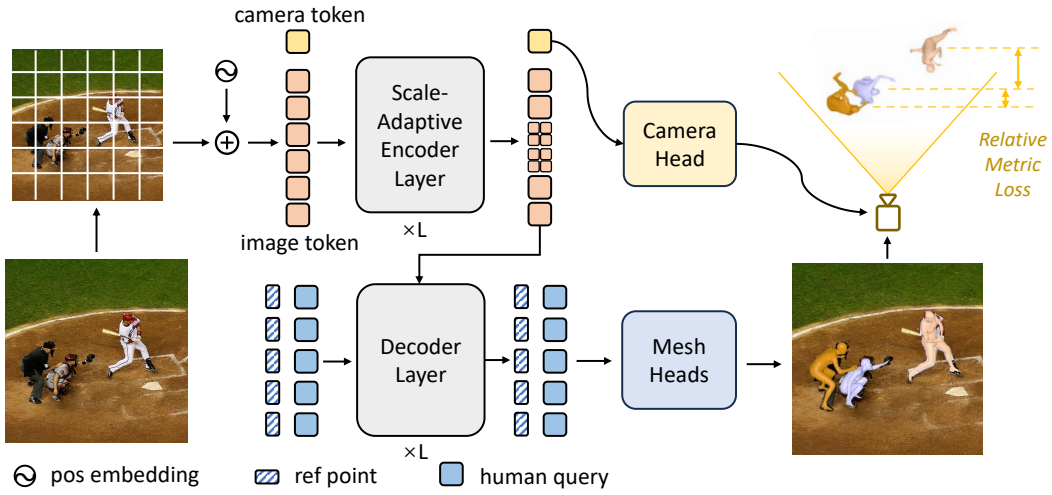

Figure 3. Architecture of our Metric-Aware HMR. We enhance a SAT-HMR backbone with two key innovations for true metric-scale recovery: a camera branch that predicts the camera's Field of View from global features using a dedicated camera token, and a relative metric loss that directly supervises the real-world distances between predicted individuals.

Once the optimal solution (s_d^*, t_d^*) are found, we compute the final depth translation $z_i^* = s_d^* d_i + t_d^*$ for each person and refine their shape parameters to match the corrected heights. This process effectively grounds the entire crowd in a physically plausible metric space.

3.2.4. DTO-Humans

To construct a large-scale in-the-wild pGT dataset with high scene-consistency, we apply DTO to CameraHMR's [41] version of 4D-Humans [11] dataset, which includes 2M images from diverse in-the-wild datasets such as InstaVariety [21], COCO [32], MPII [1], and AI Challenger [54]. To ensure the final quality and physical plausibility of our dataset, we introduce a filtering step based on the posterior probability from our optimization. For each optimized person, we calculate their standardized residual height, i.e., $(|h_i^* - \mu_i|)/\sigma_i$. We then discard any image where the average of these residuals across all individuals exceeds a threshold of 1.5. This rigorous process filters out scenes where the optimization failed to find a solution that strongly aligns with our anthropometric priors. The resulting dataset, named **DTO-Humans**, comprises 0.56 million high-quality images and 2.7 million scene-consistent person instances, featuring dense scenes with an average of 4.8 persons per image. Further details on dataset statistics and samples are provided in the supplementary material.

3.3. Metric-Aware HMR

The camera-space scene-consistent pGT generated by DTO enables us to train an end-to-end network capable of real metric-scale recovery. Thus, we propose Metric-Aware HMR (MA-HMR), a novel network that directly predicts metrically accurate meshes for all subjects in the scene.

3.3.1. Network Architecture

Our network architecture builds upon SAT-HMR [45], inheriting its Scale-Adaptive Encoder for efficient multi-scale feature extraction. As shown in Fig. 3, our key architectural modification is the introduction of a dedicated camera branch. We replace the standard classification token in the ViT-style encoder, which is designed to aggregate global scene-level information, with a learnable camera token. After passing through the encoder, the output embedding of the camera token is fed into an MLP head that regresses the camera's vertical FoV. By predicting the FoV concurrently with all human individuals' mesh parameters, the network learns the intricate coupling between global scene-level human relationship and camera properties.

3.3.2. Training Objective

FoV Loss ($\mathcal{L}_{\text{fov}}$). To supervise the camera branch, we adopt the FoV loss from HumanFoV [41], which penalizes overestimation of FoV more heavily than underestimation:

$$\mathcal{L}_{\text{fov}}(v_{\text{pred}}, v_{\text{gt}}) = \begin{cases} 3\|v_{\text{gt}} - v_{\text{pred}}\|^2 & \text{if } v_{\text{pred}} > v_{\text{gt}} \\ \|v_{\text{gt}} - v_{\text{pred}}\|^2 & \text{if } v_{\text{pred}} \leq v_{\text{gt}} \end{cases} \quad (4)$$

Relative Metric Loss ($\mathcal{L}_{\text{rm}}$). To explicitly teach the network about real-world scale, we introduce a relative metric loss to enforce plausible metric relationships between individuals. This loss operates on pairs of individuals (i, j) who are in close proximity, defined as those whose ground-truth relative depth distance is below threshold τ ($|z_{gt,i} - z_{gt,j}| < \tau$, with $\tau = 1\text{m}$). For each qualifying pair, the loss is:

$$\mathcal{L}_{\text{rm}} = \log(1 + |\text{rd}_{\text{pred}} - \text{rd}_{\text{gt}}|) \quad (5)$$

Table 1. Comparison with SOTA methods on Relative Human [48] and MuPoTS [37].

Method	f.t.	Relative Human						MuPoTS		
		$PCDR_{0.2}$↑					PCK↑	PCK(All)↑	PCK(Match)↑	MPJPE↓
		baby	kid	teen	adult	all				
CRMH [18]	✓	34.74	48.37	59.11	55.47	54.83	0.781	/	/	/
ROMP [47]	✓	30.08	48.41	51.12	55.34	54.81	0.866	63.0	67.5	127.8
BEV [48]	✓	60.77	67.09	66.07	69.71	68.27	0.884	74.7	78.2	109.0
PSVT [44]	✓	64.00	71.29	70.45	71.95	71.23	/	/	/	/
InstaHMR [31]	✓	66.67	70.85	72.19	73.95	72.86	/	73.7	76.3	/
Multi-HMR [2]	✓	/	/	/	/	/	/	84.3	89.5	90.5
SAT-HMR [45]	✓	60.15	73.95	**74.74**	72.52	72.07	0.908	**87.8**	88.8	92.2
CHMR [41]		31.25	49.90	60.07	61.20	60.43	0.921	87.3	89.0	89.5
CHMR + **DTO**		**69.65**	**76.28**	74.53	**74.41**	**74.16**	**0.936**	85.0	**90.8**	**86.3**

where rd_{gt} and rd_{pred} are the ground-truth and predicted relative depth distances between the individuals' root joints. The logarithmic form provides stable gradients and focuses on refining close-range predictions.

These two novel losses are integrated into a comprehensive objective function alongside the standard losses from the SAT-HMR framework [45]. These include the scale map loss ($\mathcal{L}_{\text{map}}$), depth loss ($\mathcal{L}_{\text{depth}}$), SMPL pose ($\mathcal{L}_{\text{pose}}$) and shape ($\mathcal{L}_{\text{shape}}$) parameter losses, a 3D joint loss ($\mathcal{L}_{\text{j3d}}$), a 2D joint reprojection loss ($\mathcal{L}_{\text{j2d}}$), L1 and GIoU bounding box losses ($\mathcal{L}_{\text{box}}$, $\mathcal{L}_{\text{giou}}$), and a focal detection loss ($\mathcal{L}_{\text{det}}$). The complete objective is a weighted sum of all components:

$$\mathcal{L} = \lambda_{\text{map}}\mathcal{L}_{\text{map}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{shape}}\mathcal{L}_{\text{shape}} \quad (6)$$
$$+\lambda_{\text{j3d}}\mathcal{L}_{\text{j3d}} + \lambda_{\text{j2d}}\mathcal{L}_{\text{j2d}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}}$$
$$+\lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{fov}}\mathcal{L}_{\text{fov}} + \lambda_{\text{rm}}\mathcal{L}_{\text{rm}}$$

where the λ terms are the corresponding loss weights.

4. Experiments

4.1. Implementation Details

For DTO, We set Detectron2's [55] detection threshold at 0.5 and nms threshold at 0.7. We discard humans whose visible keypoints predicted by ViTPose-Base [57] are less than five. We use Depth-Anything-v2-Large [59] for predicting the relative depth map and Mivolo-v2 [27] to predict human age and gender based on their body and face crops.

For MA-HMR, we adopt ViT-Base [9, 40] as the backbone. We initialize our model with the SAT-HMR's [45] stage1 checkpoint pretrained on AGORA [42], BEDLAM [3], COCO [32], MPII [1], CrowdPose [29], and Human3.6M [16], and continue training on AGORA, BEDLAM and CameraHMR's [41] version of 4D-Humans dataset [1, 21, 32, 54] for 5 epochs with denoising strategy [28, 45]. In the second stage, we add camera branch and continue training for 5 epochs with FoV loss. For pGTs from 4D-Humans, we follow SAT-HMR stage1, only to supervise the projected 2D keypoints. In the third stage, we train the model on AGORA, BEDLAM and DTO-Humans for 5 epochs with full loss. In the final stage, we finetune the model for each benchmark following [2, 45, 48]. For all training stages, we use the AdamW optimizer [36] with a learning rate of 1e-5 and weight decay of 1e-4. Data augmentation techniques include random rotation, horizontal flipping, scaling, and cropping. Training is conducted on two NVIDIA A800 GPUs with a total batch size of 64.

During evaluation of MA-HMR, we filter out low-confidence detections using a threshold of 0.3. Evaluation metrics on 3DPW [50], Hi4D [60], CMU Panoptic [19] and MuPoTS [37] include Mean Per-Joint Position Error (MPJPE), Procrustes Aligned MPJPE (PA-MPJPE), and Per-Vertex Error (PVE), all reported in millimeters (mm). The Percentage of Correct Depth Relations ($PCDR_{0.2}$) and the Percentage of Correctly estimated 2D Keypoints (PCK) using a threshold of 1/7 of human's height are evaluated on Relative Human following [48]. The Percentage of Correctly estimated 3D Keypoints (PCK) using a threshold of 15 cm is evaluated on MuPoTS. Height error (mm) is evaluated on 3DPW and Hi4D by comparing the T-pose height from the predicted shape parameters to the ground-truth.

4.2. Quantitative Comparison of DTO

To quantify the contribution of our DTO framework, we conduct an experiment to apply it as a post-processing refinement step to a state-of-the-art single-person HMR model CameraHMR [41] (CHMR). As shown in Table 1, we compare the output of CHMR and the output after refinement by our DTO framework with SOTA methods. It is important to note that neither of CHMR and DTO is fine-tuned (f.t.) on the benchmark-specific training sets, allowing for the evaluation of DTO's generalization.

On the Relative Human dataset, which is challenging due

Table 2. Comparison with SOTA methods on 3DPW [50], CMU Panoptic [19] and MuPoTS [37].

Method	f.t.	3DPW			CMU Panoptic (MPJPE↓)					MuPoTS (PCK↑)	
		MPJPE↓	PA-MPJPE↓	MVE↓	haggling	mafia	ultimatum	pizza	mean	All	Match
CMRH [18]	✓	/	/	/	129.6	133.5	153.0	156.7	143.2	/	/
3DCrowdNet [7]	✓	81.7	51.5	98.3	109.6	135.9	129.8	135.6	127.6	72.7	73.3
ROMP [47]	✓	76.6	47.3	93.4	110.8	135.9	129.8	135.6	127.6	63.0	67.5
BEV [48]	✓	78.5	46.9	92.3	90.7	103.7	113.1	125.2	109.5	74.7	78.2
PSVT [44]	✓	75.5	45.7	84.9	88.7	97.9	115.2	121.1	105.7	/	/
Insta-HMR [31]	✓	77.7	46.3	87.0	84.8	99.8	102.6	121.7	104.5	73.7	76.3
Multi-HMR [2]	✓	61.4	41.7	75.9	/	/	/	/	96.5	84.3	89.5
SAT-HMR [45]	✓	63.6	41.6	73.7	67.9	78.5	95.8	94.6	84.2	87.8	88.8
SAT-HMR stage1		81.0	52.7	94.5	101.2	120.5	135.4	120.6	119.3	83.2	86.6
SAT-HMR w. 4D		68.8	44.5	80.5	80.5	102.2	113.6	96.7	98.5	85.4	89.0
MA-HMR w. 4D		65.4	42.6	76.9	79.0	98.6	114.8	98.6	97.5	83.0	89.5
SAT-HMR **w. DH**		63.6	40.5	74.4	**64.3**	78.6	87.7	**85.5**	**79.6**	86.9	**90.0**
MA-HMR w. DH		**62.9**	**40.1**	**73.5**	66.2	**76.7**	**87.0**	87.4	**79.6**	**87.0**	89.8
SAT-HMR **w. DH**	✓	60.1	37.7	69.7	64.5	72.7	88.7	**83.6**	76.9	/	/
MA-HMR w. DH	✓	**58.5**	**36.3**	**67.9**	**62.3**	**72.3**	**84.3**	85.4	**76.3**	/	/

to its diverse ages and complex relative depths, the improvement of DTO is remarkable. The baseline CHMR, processing each person in isolation and neglecting the modeling of kids, achieves a $PCDR_{0.2}$ (all) of only 60.43. By simply applying our DTO to the baseline, the score dramatically increases by 13.73 points to 74.16, surpassing all fine-tuned SOTA methods, including the previous leader InstaHMR (+1.3). The segmentation-enhanced mesh initialization also contributes to a top-performing PCK score of 0.936, indicating superior 2D alignment.

The evaluation on the MuPoTS dataset further validates our approach. Our CHMR + DTO pipeline achieves lowest MPJPE among all competing methods. This result proves that even in less crowded scenes (with 2 or 3 adults), DTO can also solve for a consistent scale that results in more plausible human heights and thus leads to superior 3D human joints accuracy. These experiments confirm that DTO is a powerful and robust framework for enhancing scene-level consistency and metric accuracy.

4.3. Quantitative Comparison of MA-HMR

We first evaluate our model's understanding of scene-level relative relationships on the Relative Human dataset. As shown in Table 3, our fine-tuned MA-HMR sets a new SOTA with a $PCDR_{0.2}$ (all) of 75.35, surpassing both the SAT-HMR baseline (72.89) and our optimization-based CHMR + DTO (74.16). This demonstrates the advantage of end-to-end fine-tuning, allowing the model to specialize in the dataset's unique age distributions and complex relative depth cues. Notably, even without fine-tuning, our MA-HMR trained on DTO-Humans (72.90) is already on par with the fine-tuned SAT-HMR (72.89). This highlights the strong generalization provided by our model architecture when trained on our high-quality dataset.

Table 3. Results comparison on Relative Human [48].

Method	f.t.	$PCDR_{0.2}$↑					PCK↑
		baby	kid	teen	adult	all	
CHMR + **DTO**		**69.65**	**76.28**	74.53	74.41	74.16	**0.936**
SAT-HMR **w. DH**		59.14	68.98	66.20	68.83	68.30	0.927
MA-HMR w. DH		58.64	72.35	**76.83**	73.43	72.90	0.925
SAT-HMR **w. DH**	✓	65.56	74.24	73.85	73.31	72.89	0.924
MA-HMR w. DH	✓	67.76	75.57	76.81	**75.77**	**75.35**	0.926

We then evaluate human pose and shape accuracy on the classic benchmarks 3DPW, CMU Panoptic and MuPoTS with results in Table 2. The superior generalization of our dataset is evident: without benchmark-specific fine-tuning, SAT-HMR [45] trained on DTO-Humans (w. DH) significantly outperforms the same model trained on CameraHMR's version of 4D-Humans (w. 4D). This is especially evident on CMU Panoptic, where the mean MPJPE drops from 98.5 mm to 79.6 mm. Our MA-HMR architecture (MA-HMR w. DH) further improves upon this, achieving 62.9 mm on 3DPW. After fine-tuning, MA-HMR sets a new state-of-the-art on 3DPW, achieving 58.5 mm MPJPE and 36.3 mm PA-MPJPE. It also records the lowest mean error on CMU Panoptic at 76.3 mm.

On the Hi4D dataset, which focuses on close human-human interaction, our method again shows superior performance (Table 4). Simply training the baseline model on DTO-Humans (SAT-HMR w. DH) improves MPJPE from

Table 4. Comparison with SOTA methods on Hi4D [60].

Method	f.t.	MPJPE↓	PA-MPJPE↓	MVE↓
CLIFF [30]	✓	91.3	53.6	109.6
4D-Humans [11]	✓	72.1	52.4	88.6
BEV [48]	✓	91.8	52.2	101.2
TRACE [49]	✓	83.8	60.4	/
GroupRec [13]	✓	82.4	51.6	88.6
BUDDI [39]	✓	96.8	70.6	116.0
CloseInt [14]	✓	63.1	47.5	76.4
Fang et al. [10]	✓	75.0	59.7	/
ReconClose [15]	✓	59.1	44.3	72.0
SAT-HMR stage1		95.5	62.0	111.7
SAT-HMR w. 4D		74.0	50.0	90.5
MA-HMR w. 4D		68.3	47.3	85.1
SAT-HMR **w. DH**		61.0	46.4	76.9
MA-HMR w. DH		**60.2**	**45.6**	**76.0**
SAT-HMR **w. DH**	✓	44.3	34.3	54.9
MA-HMR w. DH	✓	**43.5**	**33.8**	**54.2**

Table 5. Comparison of height error (mm)↓ on 3DPW and Hi4D.

Method	3DPW	Hi4D
SAT-HMR w. 4D	49.0	111.1
MA-HMR w. 4D	41.5	86.9
SAT-HMR w. DH	38.2	38.3
MA-HMR w. DH	**36.9**	**35.0**

74.0 mm (SAT-HMR w. 4D) to 61.0 mm. After fine-tuning, our MA-HMR w. DH model reaches 43.5 mm MPJPE and 33.8 mm PA-MPJPE, outperforming all previous methods.

We validate the metric accuracy on human scale which contributes to the per-joint accuracy (Table 5). The reduction in height error when switching from w. 4D to w. DH (models here are not benchmark-specific finetuned) confirms that our dataset provides a more reliable height supervision. Furthermore, MA-HMR achieves the lowest height error, demonstrating that the relative metric loss effectively guides the network to learn plausible human scales.

4.4. Ablation Study

Ablation on DTO Components As detailed in Table 6, we evaluate the contribution of each component in our DTO framework on the Relative Human dataset, starting from the baseline CHMR model. By incorporating segmentation enhanced initial per-person estimation (+S), we slightly improve 2D pose estimation (PCK 0.921 → 0.936) as the model can better focus on the target human. By leveraging Inter-human Depth relation for translation optimization (+D), the $\text{PCDR}_{0.2}$ (all) score jumps from 60.89 to 72.15. This demonstrates that optimizing for relative depth is the most critical factor in resolving scene-wide layout. Finally,

Table 6. Ablation studies on DTO on Relative Human. S: segmentation-enhanced inference. D: optimization based on inter-human depth relation. X: intra-human depth scale constraints.

Method	$\text{PCDR}_{0.2}$↑					PCK↑
	baby	kid	teen	adult	all	
CHMR	31.25	49.90	60.07	61.20	60.43	0.921
CHMR+S	30.80	50.93	62.41	61.29	60.89	0.936
CHMR+S+D	65.77	72.40	74.53	72.38	72.15	0.936
CHMR+S+D+X	**69.65**	**76.28**	**74.53**	**74.41**	**74.16**	**0.936**

Table 7. Ablation studies on MA-HMR.

camera	$\mathcal{L}_{\text{rm}}$	3DPW		Hi4D	
		MPJPE↓	MVE↓	MPJPE↓	MVE↓
		63.6	74.4	61.0	76.9
✓		63.1	74	61.7	77.4
✓	✓	**62.9**	**73.5**	**60.2**	**76.0**

we add the intra-human depth scale constraint (+X). This further improves the $\text{PCDR}_{0.2}$ (all) score from 72.15 to 74.16. This shows that enforcing this scale constraint prevents the optimization from settling on physically unrealistic scale factors and achieve a more robust estimation.

Ablation on MA-HMR Architecture In Table 7, we analyze our architectural choices for MA-HMR. All models are trained on DTO-Humans without benchmark-specific finetuning. We find that adding only the camera branch improves performance on 3DPW but slightly degrades on Hi4D (MPJPE 61.0 → 61.7). We attribute this to Hi4D's uniform camera FoV settings ($59.53° \pm 0.07°$), which align with the baseline's fixed $60°$ default. The unconstrained FoV prediction introduces unnecessary variance. However, combining the camera branch with our relative metric loss ($\mathcal{L}_{\text{rm}}$) yields the best performance across all benchmarks. This confirms that $\mathcal{L}_{\text{rm}}$ provides essential supervision, enabling the network to effectively resolve the ambiguity between camera intrinsics and human scale.

5. Conclusion

In this paper, we introduce DTO, an optimization framework for scene-consistent pGT generation, resulting in the DTO-Humans dataset. We also propose MA-HMR, an end-to-end network featuring a camera branch and a relative metric loss to achieve metric-scale recovery. Extensive experiments demonstrate that our method achieves SOTA performance in human relative depth reasoning and pose and shape accuracy, confirming that our DTO-Humans dataset and MA-HMR effectively bridge the gap between relative scene understanding and absolute metric reconstruction.

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. 2, 3, 5, 6

[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, 7, 8

[3] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 6

[4] 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

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[6] Centers for Disease Control and Prevention. Growth charts - percentile data files with lms values, 2024. 2, 3, 4

[7] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1475–1484, 2022. 7

[8] 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. 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[10] Qi Fang, Yinghui Fan, Yanjun Li, Junting Dong, Dingwei Wu, Weidong Zhang, and Kang Chen. Capturing closely interacted two-person motions with reaction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 655–665, 2024. 8

[11] 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, 5, 8, 6

[12] 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

[13] Buzhen Huang, Jingyi Ju, Zhihao Li, and Yangang Wang. Reconstructing groups of people with hypergraph relational reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14873–14883, 2023. 8

[14] 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. 8

[15] 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. 8

[16] 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. 6

[17] Aline Jelenkovic, Yoon-Mi Hur, Reijo Sund, Yoshie Yokoyama, Sisira H Siribaddana, Matthew Hotopf, Athula Sumathipala, Fruhling Rijsdijk, Qihua Tan, Dongfeng Zhang, et al. Genetic and environmental influences on adult human height across birth cohorts from 1886 to 1994. *Elife*, 5:e20320, 2016. 2, 3, 4

[18] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2020. 3, 6, 7

[19] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. 2, 6, 7

[20] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 2, 3

[21] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 2, 5, 6

[22] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1715–1725, 2022. 3

[23] 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

[24] 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

[25] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11605–11614, 2021. 3

[26] Maksim Kuprashevich and Irina Tolstykh. Mivolo: Multi-input transformer for age and gender estimation. In *International conference on analysis of images, social networks and texts*, pages 212–226. Springer, 2023. 3

[27] Maksim Kuprashevich, Grigorii Alekseenko, and Irina Tolstykh. Beyond specialization: assessing the capabilities of mllms in age and gender estimation. *arXiv preprint arXiv:2403.02302*, 2024. 2, 3, 4, 6

[28] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022. 3, 6

[29] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 6

[30] 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, 8

[31] Xinyao Liao, Chen Zhang, Jianyao Xu, Wanjuan Su, Zhi Chen, and Wenbing Tao. Instahmr: Instance-aware one-stage multi-person human mesh recovery. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 6, 7

[32] 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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 2, 3, 5, 6

[33] Jingyuan Liu, Nazmus Saquib, Chen Zhutian, Rubaiat Habib Kazi, Li-Yi Wei, Hongbo Fu, and Chiew-Lan Tai. Posecoach: A customizable analysis and visualization system for video-based running coaching. *IEEE Transactions on Visualization and Computer Graphics*, 30(7):3180–3195, 2022. 2

[34] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 3

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. 3

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[37] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 international conference on 3D vision (3DV)*, pages 120–130. IEEE, 2018. 2, 6, 7

[38] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2307, 2022. 2, 3

[39] Lea Müller, 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. 8, 6

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[41] 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, 3, 5, 6

[42] 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. 3, 6

[43] 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

[44] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21254–21263, 2023. 2, 3, 6, 7

[45] 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 Computer Vision and Pattern Recognition Conference (CVPR)*, pages 16796–16806, 2025. 5, 6, 7, 9

[46] 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, pages 1834–1843, 2024. 2, 3

[47] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021. 3, 6, 7

[48] 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, 6, 7, 8

[49] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8866, 2023. 8

[50] 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. 2, 6, 7

[51] Jingbo Wang, Ye Yuan, Zhengyi Luo, Kevin Xie, Dahua Lin, Umar Iqbal, Sanja Fidler, and Sameh Khamis. Learning human dynamics in autonomous driving scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20796–20806, 2023. 2

[52] 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 Computer Vision and Pattern Recognition Conference*, pages 21991–22000, 2025. 3

[53] 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. 3

[54] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 2, 5, 6

[55] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3, 6

[56] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2

[57] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 3, 6

[58] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 3

[59] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 2, 3, 4, 6

[60] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose 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*, pages 17016–17027, 2023. 2, 6, 8

[61] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 546–556, 2021. 3

Towards Metric-Aware Multi-Person Mesh Recovery by Jointly Optimizing Human Crowd in Camera Space

Supplementary Material

6. DTO

6.1. Analytical Solution to the DTO Problem

We aim to solve the following constrained quadratic minimization problem, where the scale parameter s_d is bounded by a plausible range derived from the intra-human depth scale:

$$\min_{s_d,t_d} \sum_{i=1}^{K} \frac{\left(\hat{h}_i \cdot \frac{s_d d_i + t_d}{\hat{z}i} - \mu_i\right)^2}{\sigma_i^2} \quad \text{s.t.} \quad X_{\min} \le s_d \le X_{\max} \tag{7}$$

Simplification of Terms To simplify the derivation, we define $a_i = \frac{\hat{h}_i d_i}{\hat{z}_i \sigma_i}$, $b_i = \frac{\hat{h}_i}{\hat{z}_i \sigma_i}$, and $c_i = \frac{\mu_i}{\sigma_i}$. The objective function, which we denote as $L(s_d, t_d)$, can now be rewritten as a simple sum of squares:

$$L(s_d, t_d) = \sum_{i=1}^{K} (a_i s_d + b_i t_d - c_i)^2 \tag{8}$$

The problem has two inequality constraints, which we write in the standard form $g(x) \le 0$:

$$g_1(s_d, t_d) = X_{\min} - s_d \le 0 \tag{9}$$
$$g_2(s_d, t_d) = s_d - X_{\max} \le 0 \tag{10}$$

The Lagrangian and KKT Conditions The Lagrangian $\mathcal{L}_{\text{g}}$ for this problem involves two Lagrange multipliers, λ_1 and λ_2, corresponding to the two constraints:

$$\mathcal{L}_{\text{g}}(s_d, t_d, \lambda_1, \lambda_2) = L(s_d, t_d) + \lambda_1(X_{\min} - s_d) + \lambda_2(s_d - X_{\max}) \tag{11}$$

The KKT conditions for optimality are:

$$\begin{cases} \dfrac{\partial \mathcal{L}_{\text{g}}}{\partial s_d} = \displaystyle\sum_{i=1}^{K} 2a_i(a_i s_d + b_i t_d - c_i) - \lambda_1 + \lambda_2 = 0, & \text{(12a)} \\ \dfrac{\partial \mathcal{L}_{\text{g}}}{\partial t_d} = \displaystyle\sum_{i=1}^{K} 2b_i(a_i s_d + b_i t_d - c_i) = 0, & \text{(12b)} \\ X_{\min} - s_d \le 0, & \text{(12c)} \\ s_d - X_{\max} \le 0, & \text{(12d)} \\ \lambda_1, \lambda_2 \ge 0, & \text{(12e)} \\ \lambda_1(X_{\min} - s_d) = 0, & \text{(12f)} \\ \lambda_2(s_d - X_{\max}) = 0 & \text{(12g)} \end{cases}$$

Solving the System The nature of the convex quadratic objective and the linear constraints allows for a straightforward solution strategy. The approach is to first solve the unconstrained minimum of the objective function, assuming both constraints are inactive ($\lambda_1 = \lambda_2 = 0$). In this scenario, the stationarity conditions (Eqs. (12a) and (12b)) simplify to a standard 2x2 linear system:

$$\begin{cases} \left(\sum a_i^2\right) s_d + \left(\sum a_i b_i\right) t_d = \sum a_i c_i & \text{(13a)} \\ \left(\sum a_i b_i\right) s_d + \left(\sum b_i^2\right) t_d = \sum b_i c_i & \text{(13b)} \end{cases}$$

Solving this system yields the unconstrained solution, denoted $(s_{d,\text{unc}}, t_{d,\text{unc}})$. We now check where $s_{d,\text{unc}}$ falls relative to the interval $[X_{\min}, X_{\max}]$. There are three possible outcomes:

Case 1: The unconstrained solution is feasible. If $X_{\min} \le s_{d,\text{unc}} \le X_{\max}$, the unconstrained minimum already satisfies the constraints. In this case, the KKT conditions with $\lambda_1 = \lambda_2 = 0$ are fully satisfied. The global optimum is the unconstrained solution:

$$(s_d^*, t_d^*) = (s_{d,\text{unc}}, t_{d,\text{unc}}) \tag{14}$$

Case 2: The unconstrained solution is below the lower bound. If $s_{d,\text{unc}} < X_{\min}$, the convexity of the objective function implies that the minimum over the feasible set must lie on the boundary closest to the unconstrained minimum. Therefore, the optimal scale s_d^* is clamped to the lower bound:

$$s_d^* = X_{\min} \tag{15}$$

This corresponds to the KKT case where the lower bound is active ($\lambda_1 > 0$) and the upper bound is inactive ($\lambda_2 = 0$). We substitute $s_d^* = X_{\min}$ into the second stationarity condition (Eq. (12b)), which is independent of the multipliers, to solve for the optimal t_d^*:

$$\left(\sum b_i^2\right) t_d^* = \sum (b_i c_i - a_i b_i X_{\min}) \tag{16}$$

$$t_d^* = \frac{\sum b_i c_i - X_{\min} \sum a_i b_i}{\sum b_i^2} \tag{17}$$

Case 3: The unconstrained solution is above the upper bound. If $s_{d,\text{unc}} > X_{\max}$, by the same logic, the solution is clamped to the upper bound:

$$s_d^* = X_{\max} \tag{18}$$

This corresponds to the KKT case where the upper bound is active ($\lambda_2 > 0$) and the lower bound is inactive ($\lambda_1 = 0$). We substitute $s_d^* = X_{\max}$ into Eq. (12b) to find the corresponding t_d^*:

$$t_d^* = \frac{\sum b_i c_i - X_{\max} \sum a_i b_i}{\sum b_i^2} \tag{19}$$

This procedure of solving the unconstrained problem and then clamping the solution to the feasible interval $[X_{\min}, X_{\max}]$ is guaranteed to find the unique global minimum of this convex, constrained quadratic program.

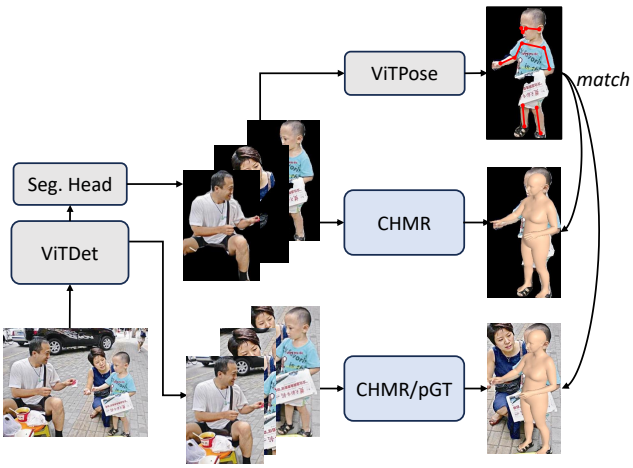

Figure 4. Initial Per-Person Estimation Pipeline.

6.2. Initial Per-Person Estimation

To leverage 2D landmarks to filter out inaccurate mesh estimations, we introduce a two-stage matching process. The 2D keypoints detected by ViTPose on the masked crops serve as the *query* set for this process. Specifically, an instance is considered a valid match if over half of its visible keypoints align with projected joints. An individual keypoint-to-joint correspondence is deemed successful if their pixel distance is less than half the person's head height.

In the first stage, we match these query keypoints against a high-fidelity target. For datasets with available pGT, we use the keypoints projected from the pGT mesh as the initial target. If pGT is not available, we use the keypoints projected from an initial mesh predicted by CameraHMR on the original, unmasked crop to leverage full crop context.

In the second stage, any query keypoints that remain unmatched are subsequently compared against the keypoints projected from the mesh inferred from the masked crop. This allows us to find correspondences for bodies that might be better aligned in the segmentation-enhanced prediction.

This hierarchical strategy ensures we prioritize high-quality correspondences while still leveraging the fine-grained alignment from the masked prediction.

6.3. Intra-human Scale Factor Bounds

The hyperparameters α_1 and α_2, which define the lower and upper bounds for the global depth scale s_d, are set dynamically based on the scene's spatial layout. We first identify quasi-planar scenes by checking if the variance of inter-person depths is smaller than the average intra-person depth variance. In such cases, where all individuals are roughly equidistant from the camera, the intra-human scale X is the most reliable cue. Therefore, we set $\alpha_1 = \alpha_2 = 1$, effectively fixing the scale factor to $s_d = X$ and solving for t_d only.

Otherwise, for scenes with clear depth separation, we set $\alpha_1 = 1$ and $\alpha_2 = 5$. This wider range is motivated by the behavior of initial camera estimation pipeline (HumanFoV). Their FoV loss leads to larger predicted focal lengths to prevent artifacts, which in turn results in overestimated camera-space human depth translations (z_i). By allowing s_d to be larger than the intra-human scale X, we give the optimizer a sufficient range to find a globally consistent scale.

6.4. Height Priors

Priors for Minors For subjects under 15 years old, we categorize them into three age groups: 0–3 years (Baby), 3–8 years (Kid), and 8–15 years (Teen). The height priors are derived from demographic growth statistics from Centers for Disease Control and Prevention (CDC) [6]. All height values are in meters.

- **0–3 years (Baby):** We use the statistical height distributions for infants at 0, 3, 6, 9, 12, 18, 24, 30, and 36 months of age. By aggregating the data points sampled from these monthly distributions, we fit a single composite Gaussian, resulting in a prior of $\mathcal{N}(0.801, 0.126^2)$.
- **3–8 years (Kid):** For this age range, we treat the height distribution for each year as a distinct Gaussian. We then form a an empirical distribution from the sum of these Gaussians and approximate it with a single, unified Gaussian, yielding a prior of $\mathcal{N}(1.122, 0.120^2)$.
- **8–15 years (Teen):** Following the same procedure, we model the mixture of annual height distributions for the Teen with a single Gaussian, resulting in a prior of $\mathcal{N}(1.477, 0.156^2)$.

A visualization of the original height distribution and the final fitted Gaussian distribution for each group is in Fig. 5.

Priors for Adults For subjects over 15 years old, we adopt a hybrid approach for their height priors. Models like CHMR (as its pGT generation method CamSMPLify) often underestimate a person's height in non-standing poses (Fig. 11) in their effort to achieve a more accurate 2D keypoint reprojection. This can bias the scene's scale. To counteract this, we combine the model's prediction with a gender-specific statistical prior. Let $\hat{h}_{\text{CHMR}}$ be the initial height predicted by CHMR for an adult. We leverage demographic data [17] which models male height as

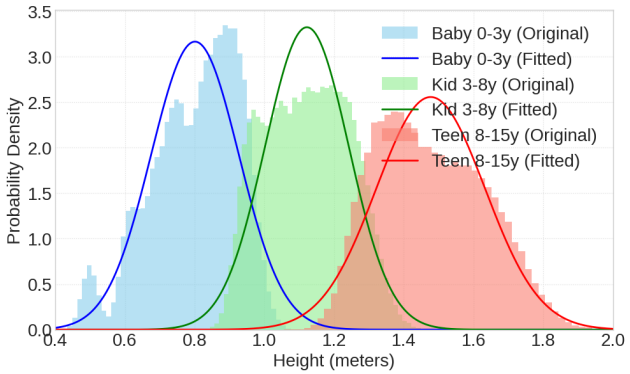

Figure 5. Height Priors for Minors.

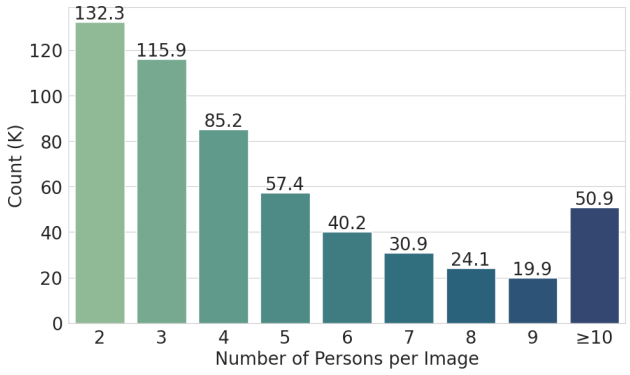

Figure 6. Number of persons per image in DTO-Humans.

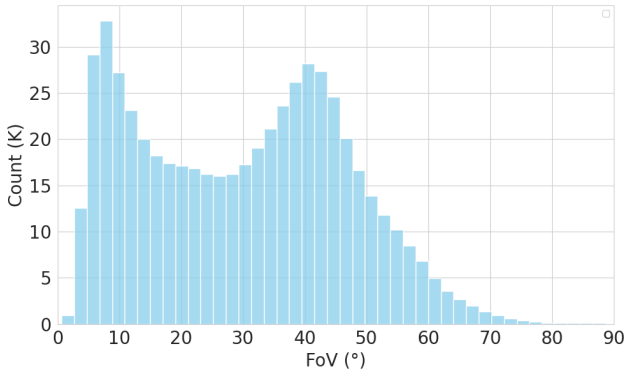

Figure 7. FoV distribution in DTO-Humans

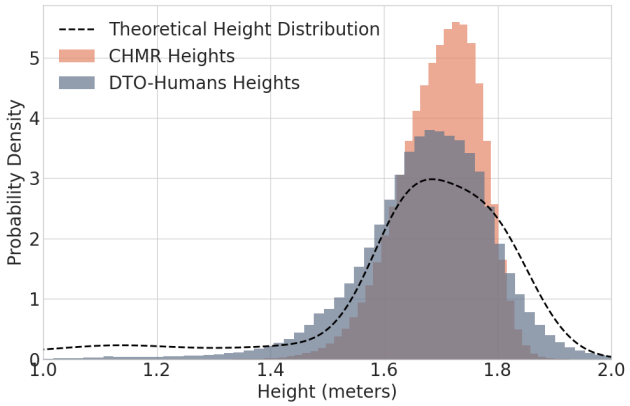

Figure 8. Comparison of Height Distributions.

$\mathcal{N}(1.784, 0.076^2)$ and female height as $\mathcal{N}(1.647, 0.071^2)$. We denote the mean and standard deviation for the person's detected gender as μ_{gender} and σ_{gender}, respectively. The final height prior for an adult is then defined as: $\mu_i = \frac{\hat{h}_{\text{CHMR}} + \mu_{\text{gender}}}{2}$, $\sigma_i^2 = \sigma_{\text{gender}}^2$. This formulation anchors the height estimate by averaging the model's prediction with a reliable demographic mean, while retaining the variance from the population statistics to account for natural diversity. The effectiveness of this hybrid prior is validated in our ablation study on the MuPoTS dataset (Sec 9).

7. DTO-Humans

The DTO-Humans dataset is curated to reflect challenging, real-world conditions. As shown in Figure 6, the dataset features a rich distribution of scenes with varying numbers of people. A large number (50.9K) of the images contain ten or more individuals, providing ample data with crowd scene consistency. Furthermore, Figure 7 illustrates that the dataset encompasses a wide distribution of camera FoV. This diversity in both population density and camera parameters makes DTO-Humans a valuable training dataset for robust 3D human scene understanding.

The height distribution of our DTO-Humans dataset demonstrates human size diversity derived from DTO framework, as illustrated in Fig. 8. A direct comparison with the CHMR's [41] version of 4D-Humans dataset reveals that the latter's height estimations are heavily concentrated in a narrow band (1.6m to 1.8m). In contrast, DTO-Humans provides a much broader and more realistic spectrum of human heights, effectively capturing individuals who are shorter or taller than this typical range.

We also analyze our dataset's distribution against a theoretical height distribution modeling a population with a uniform age distribution. While DTO-Humans aligns more closely with this ideal distribution than the baseline, the deviation caused by under-representation of non-adults remains. Specifically, minors constitute approximately 6.2% of the people in our dataset. This is lower than the 18.8% (i.e., 15/80) proportion that would be expected in a uniform demographic sample. Consequently, while DTO-Humans offers a broader representation of real-world height diversity, its composition still reflects the common challenge of sourcing images with a high prevalence of kids.

The visualization of samples in DTO-Humans is presented in Fig. 9, 10, 17, 18.

Figure 9. Visualization of samples in DTO-Humans

Figure 10. Visualization of samples in DTO-Humans

8. MA-HMR

8.1. Datasets

As mentioned in the main paper, our method is trained and evaluated on multiple publicly available datasets. Here, we provide additional details regarding their composition and usage in our experiments.

AGORA [42] is a large-scale synthetic collection built from 4,240 high-quality textured human scans (including 257 child scans) rendered in scenes containing 5-15 persons per image, resulting in approximately 14K training images and 173K individual person crops. BEDLAM [3] is likewise synthetic, featuring monocular RGB videos of full-body humans with ground-truth 3D body parameters; it covers diverse body shapes, skin tones, clothing and motion in realistic scenes. In our experiments we use the training splits of both AGORA and BEDLAM (with SMPL annotations) for training MA-HMR, leveraging the broad variation of synthetic data before fine-tuning on real-world data.

4D-Humans dataset [11] is a large-scale in-the-wild training collection that incorporates images culled from diverse sources such as InstaVariety [21], COCO [32], MPII [1] and AI Challenger [54], thereby enhancing generalization to real-world scenes. In our work, we leverage the 4D-Humans dataset to generate DTO-Humans.

Relative Human [48] is an in-the-wild multi-person dataset annotated with relative depth layers and age-group labels. Each image contains several people, and annotations include the depth-ordering of all persons: individuals whose depth difference is less than 0.3m share the same layer. In addition, age categories (adult, teenager, child, baby) are provided to help resolve height-depth ambiguity. It consists of approximately 7.6K images and over 24.8K person instances. In our work we use RH to evaluate the relative-depth accuracy of our mesh reconstruction results.

3DPW [50], CMU Panoptic [19] and MuPoTS [37] datasets are three canonical multi-person human mesh recovery benchmarks. Specifically, 3DPW provides challenging in-the-wild multi-person scenes with moving cameras. CMU Panoptic covers precisely captured multi-person interactions in a controlled studio setting. MuPoTS focuses on generalization by providing a variety of realistic scenes for evaluation. In our work we adopt the standard evaluation protocols of prior work such as Multi-HMR [2] and SAT-HMR [45] on these benchmarks.

Hi4D [60] focuses on close-interaction scenarios of two adults in prolonged physical contact. It comprises 20 subject pairs, covers 100 sequences and provides over 11K frames of 4D textured scans and SMPL annotations. As there is no official split, we follow the splitting protocol adopted in BUDDI [39]. In our work, we leverage Hi4D to further assess the performance of our model in close-interaction scenes.

8.2. Parameter Overhead

The proposed MA-HMR introduces a minimal parameter overhead compared to the baseline SAT-HMR. The addition of a 4-layer MLP for camera FoV regression slightly increases the total parameter count from 221.9M to 223.7M (+0.8%), demonstrating the efficiency of our approach.

8.3. Hyperparameters

The weighting coefficients for each term in MA-HMR's training loss function, λ_{map}, λ_{depth}, λ_{pose}, λ_{shape}, λ_{j3ds}, λ_{j2ds}, λ_{box}, λ_{det}, λ_{fov}, and λ_{rm}, are set to 4.0, 0.5, 5.0, 3.0, 8.0, 40.0, 2.0, 1.0, 0.5 and 0.5, respectively.

8.4. Qualitative Results

In Fig. 13, we compare with Multi-HMR under different FoV settings. We then compare with SAT-HMR on 3DPW and CMU Panoptic (Fig.14, 15), revealing our pose accuracy and spatial consistency. Fig. 16 presents results on Relative Human, where large variations in subject distance and scale further demonstrate the robustness of our approach.

9. More Ablations

9.1. Ablation on Gender-Aware Height Prior

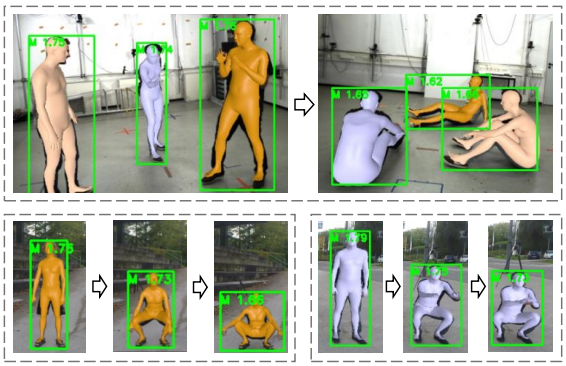

Figure 11. Illustration of height estimation bias in CameraHMR. The estimated heights are shown in green (in meters).

Table 8. Ablation study on MuPoTS evaluating our DTO components. S: segmentation-enhanced inference. D': optimization with gender-agnostic height prior. D: optimization with with gender-aware height prior. X: intra-human depth scale constraints.

Method	MPJPE ↓
CHMR	89.5
CHMR+S	87.2
CHMR+S+D'	87.1
CHMR+S+D'+X	86.7
CHMR+S+D+X	**86.3**

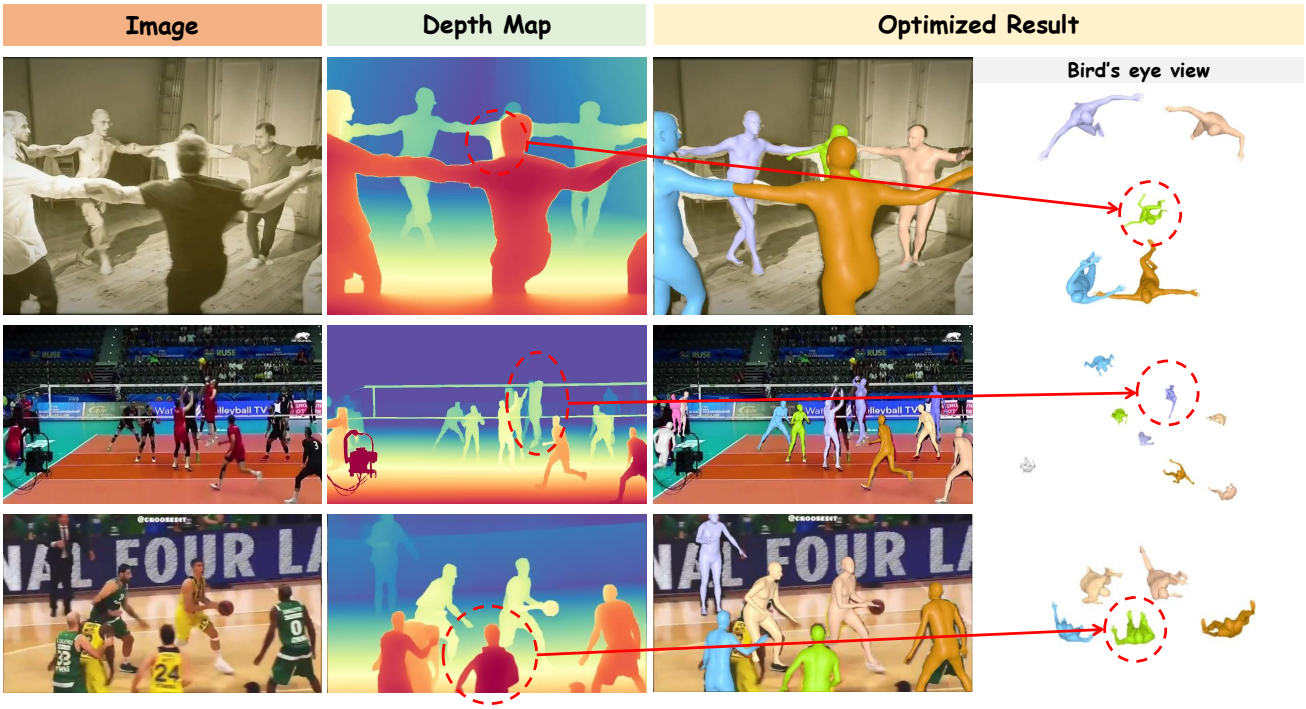

Figure 12. Examples of DTO failure cases, primarily caused by inaccurate relative depth maps.

We evaluate gender-aware height prior on MuPoTS, and the quantitative results are shown in Table 8. The baseline CHMR is progressively improved by adding segmentation-enhanced inference (+S). We then compare our full model (+D+X), which uses the gender-aware prior, against a variant where DTO is applied without this prior (+D'+X).

The introduction of the gender-aware prior (+D) further reduces the MPJPE from 86.7 to 86.3. This confirms that guiding the optimization with a statistically-grounded height mean effectively mitigates the model's underestimation bias, leading to a more accurate final 3D reconstruction.

Table 9. Performance evaluation of different Depth Anything v2 (DAv2) backbones on Relative Human.

HMR Model	Depth Model	$PCDR_{0.2}$ (all) ↑
InstaHMR [31]	/	72.83
CHMR	DAv2-S	72.93
CHMR	DAv2-B	73.74
CHMR	DAv2-L	**74.16**

9.2. Ablation on Depth Model

To select the optimal depth estimation model for our framework, we evaluated different backbones for the Depth Anything v2 model on the Relative Human dataset. As shown in Table 9, we tested ViT-Small, Base and Large.

It is noteworthy that even using the smallest DAv2-S backbone yields a PCDR score of 72.93, which surpasses the state-of-the-art performance of InstaHMR (72.86), indicating the low entry barrier and high efficiency of our DTO framework. As DAv2-L offers the best performance for our task, we selected it as our default depth model.

10. Limitations

The limitation of our DTO framework lies in its dependency on the accuracy of the upstream relative depth estimation model. Errors or ambiguities in the generated depth map can propagate, leading to implausible scene reconstructions. Fig. 12 illustrates several typical failure modes. Firstly, severe occlusion can cause the depth model to assign a foreground depth value to a background person, making DTO incorrectly scale them down (top row). In scenes with ambiguous depth cues, such as an athlete jumping mid-air, the depth model may resort to flawed heuristics (e.g., higher in image means farther away), misplacing the person in the scene (middle row). Lastly, partial visibility without a clear ground plane can confuse the depth model, leading it to incorrectly flatten the relative depths of multiple people, which DTO then inherits (bottom row). On the other hand, these cases underscore that future advancements in monocular depth estimation can directly enhance the accuracy and robustness of our DTO framework, thus lead to better 3D human scene understanding.

Figure 13. Qualitative Comparison with Multi-HMR [2]. Multi-HMR supports an optional FoV input to recover human meshes in camera space. For comparison, we test Multi-HMR in different FoV settings. (The definition of the displayed FoV here follows Multi-HMR, which is the larger of the vertical FoV and horizontal FoV). Notably, in the bottom cases, our MA-HMR successfully handles both the near-field action (the dunk over a defender) and the far-field subjects (the referee outside the three-point line and the spectators). In contrast, when we narrow Multi-HMR's input FoV for correctly placing the distant referee, its reconstruction of the near-field dunk deteriorates.

Figure 14. Qualitative Results of MA-HMR on 3DPW, with comparison to SAT-HMR [45]

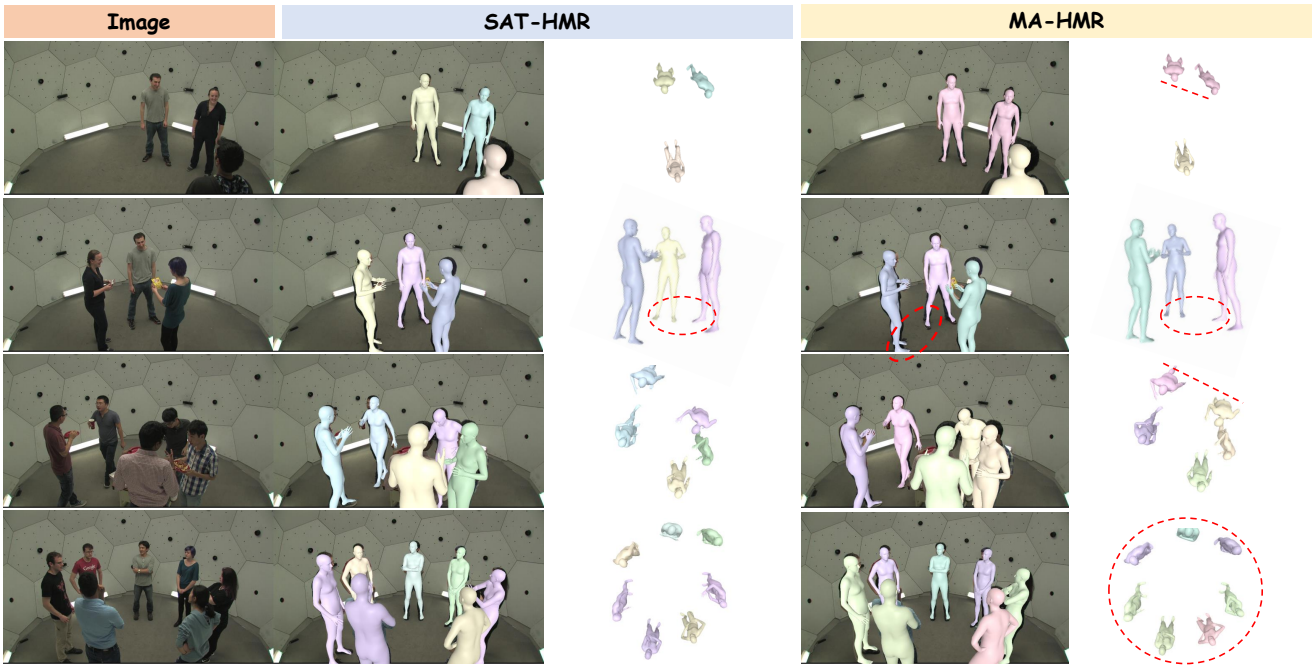

Figure 15. Qualitative Results of MA-HMR on CMU Panoptic, with comparison to SAT-HMR [45]

Figure 16. Qualitative Results of MA-HMR on Relative Human.

Figure 17. More visualization of samples in DTO-Humans.

Figure 18. More visualization of samples in DTO-Humans.