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

Supplementary Material

7. DTO

7.1. 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.

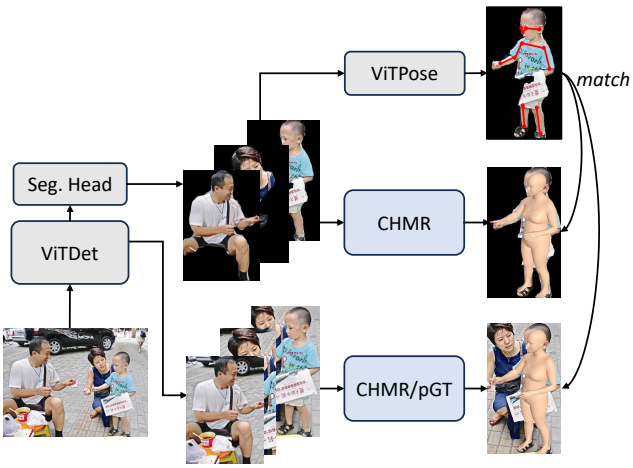

Figure 4. Initial Per-Person Estimation Pipeline.

7.2. 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.

7.3. 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.

7.4. 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) [7]. 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. 15) 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 [18] which models male height as

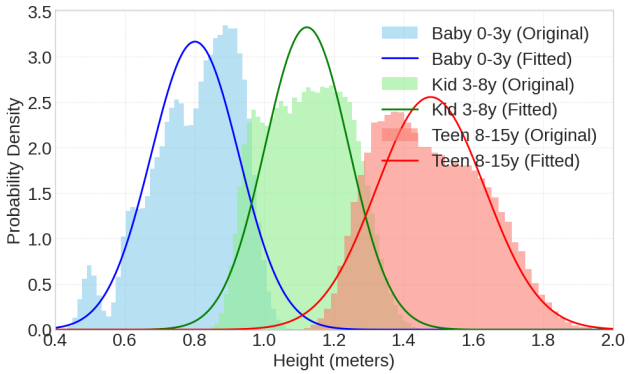

Figure 5. Height Priors for Minors.

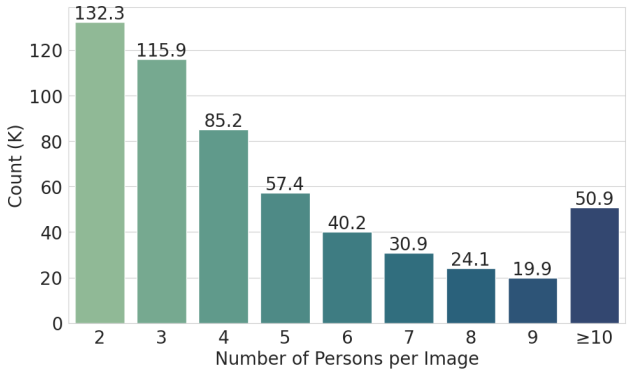

Figure 6. Number of persons per image in DTO-Humans.

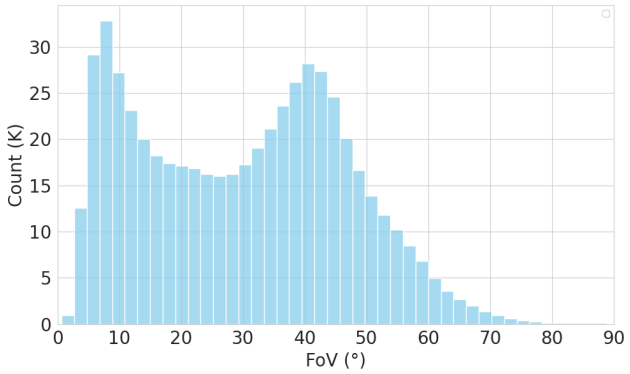

Figure 7. FoV distribution in DTO-Humans

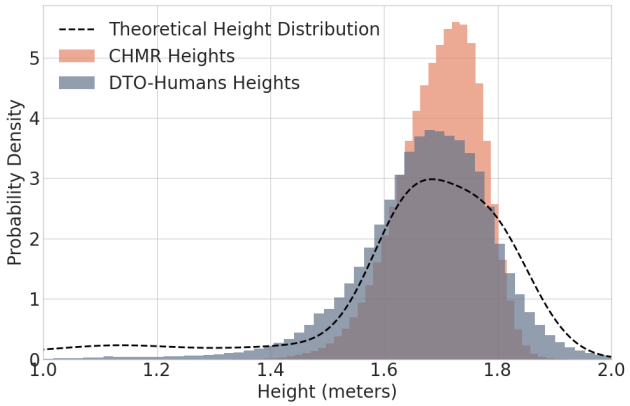

Figure 8. Comparison of Height Distributions.

$\mathcal{N}(1.784, 0.076^2)$ 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 10).

8. 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 [42] 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

9. MA-HMR

9.1. 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 [43] 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 [12] is a large-scale in-the-wild training collection that incorporates images culled from diverse sources such as InstaVariety [22], COCO [33], MPII [1] and AI Challenger [55], thereby enhancing generalization to real-world scenes. In our work, we leverage the 4D-Humans dataset to generate DTO-Humans.

Relative Human [49] 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 [51], CMU Panoptic [20] and MuPoTS [38] 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 [46] on these benchmarks.

Hi4D [61] 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 [40]. In our work, we leverage Hi4D to further assess the performance of our model in close-interaction scenes.

9.2. 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.

9.3. 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.

9.4. Qualitative Results

In Fig. 11, we compare with Multi-HMR under different FoV settings. We then compare with SAT-HMR on 3DPW and CMU Panoptic (Fig.12, 13), revealing our pose accuracy and spatial consistency. Fig. 14 presents results on Relative Human, where large variations in subject distance and scale further demonstrate the robustness of our approach.

Figure 11. 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 12. Qualitative Results of MA-HMR on 3DPW, with comparison to SAT-HMR [46]

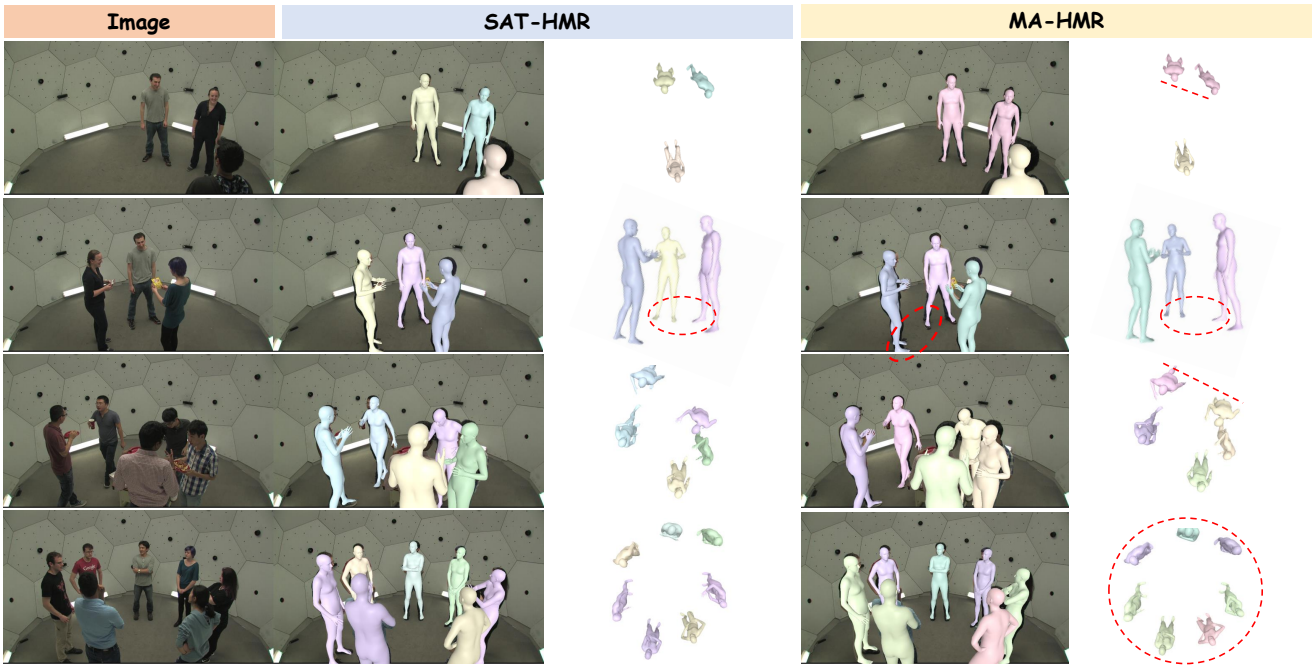

Figure 13. Qualitative Results of MA-HMR on CMU Panoptic, with comparison to SAT-HMR [46]

Figure 14. Qualitative Results of MA-HMR on Relative Human.

10. More Ablations

10.1. Ablation on Gender-Aware Height Prior

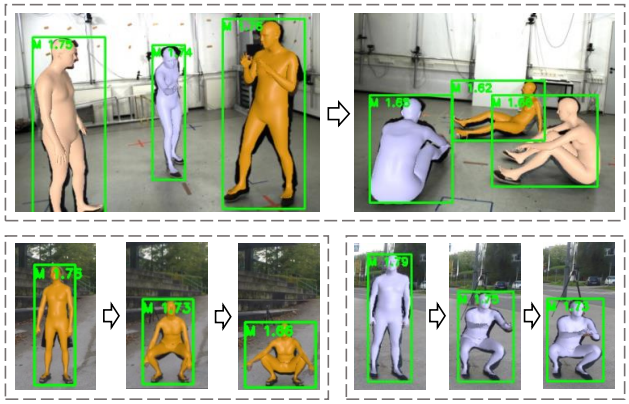

Figure 15. Illustration of height estimation bias in CameraHMR. The estimated heights are shown in green (in meters).

Table 8. Ablation study on MuPoTS evaluating our DTO components. S: segmentation-enhanced inference. D': optimization with gender-agnostic height prior. D: optimization with 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**

We evaluate the impact of the gender-aware height prior on MuPoTS, with quantitative results 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 with a gender-agnostic 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 demographic height mean effectively mitigates the model's underestimation bias, leading to a more accurate final 3D reconstruction.

10.2. Evaluation of Absolute Height Accuracy

To validate the metric accuracy on the human scale, which directly contributes to the overall per-joint accuracy, we evaluate the absolute height error on the 3DPW and Hi4D datasets (Table 9). The height error is calculated by comparing the T-pose height derived from the predicted shape parameters against the ground-truth.

Table 9. Comparison of absolute 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**

The significant reduction in height error when switching from the baseline dataset (w. 4D) to our DTO-Humans dataset (w. DH) confirms that our data generation pipeline provides much more reliable metric scale supervision. Furthermore, MA-HMR achieves the lowest overall height error, demonstrating that our proposed relative metric loss effectively guides the end-to-end network to internalize plausible human scales. Note that the models evaluated here are not benchmark-specifically fine-tuned.

10.3. Ablation on Metric Depth Estimators

We evaluate whether off-the-shelf metric depth estimators can directly replace our DTO framework for resolving human-scale ambiguity. We substituted our relative depth and height prior pipeline with state-of-the-art metric depth models, specifically the metric version of Depth Anything v2 (DAv2-m) and Depth Pro [4]. As shown in Table 10, using these direct metric depth estimators significantly degrades performance on the MuPoTS dataset.

The MPJPE increases from the baseline 89.5 mm to 113.6 mm (Depth Pro) and over 124.7 mm (DAv2-m), frequently resulting in implausible absolute human heights (e.g., adults shorter than 1.2m or taller than 2m). This confirms that general-purpose metric depth estimators are currently unreliable for fine-grained human-scale recovery. In contrast, our DTO framework—which anchors relative depth using robust statistical height priors—provides a significantly more stable and accurate metric reconstruction for human bodies (lowering MPJPE to 86.3 mm).

Table 10. Ablation study on MuPoTS evaluating metric depth estimators.

Method	MPJPE (↓)
CHMR	89.5
+ DAv2-m (indoor)	124.7
+ DAv2-m (outdoor)	150.5
+ Depth Pro	113.6
CHMR + DTO (Ours)	**86.3**

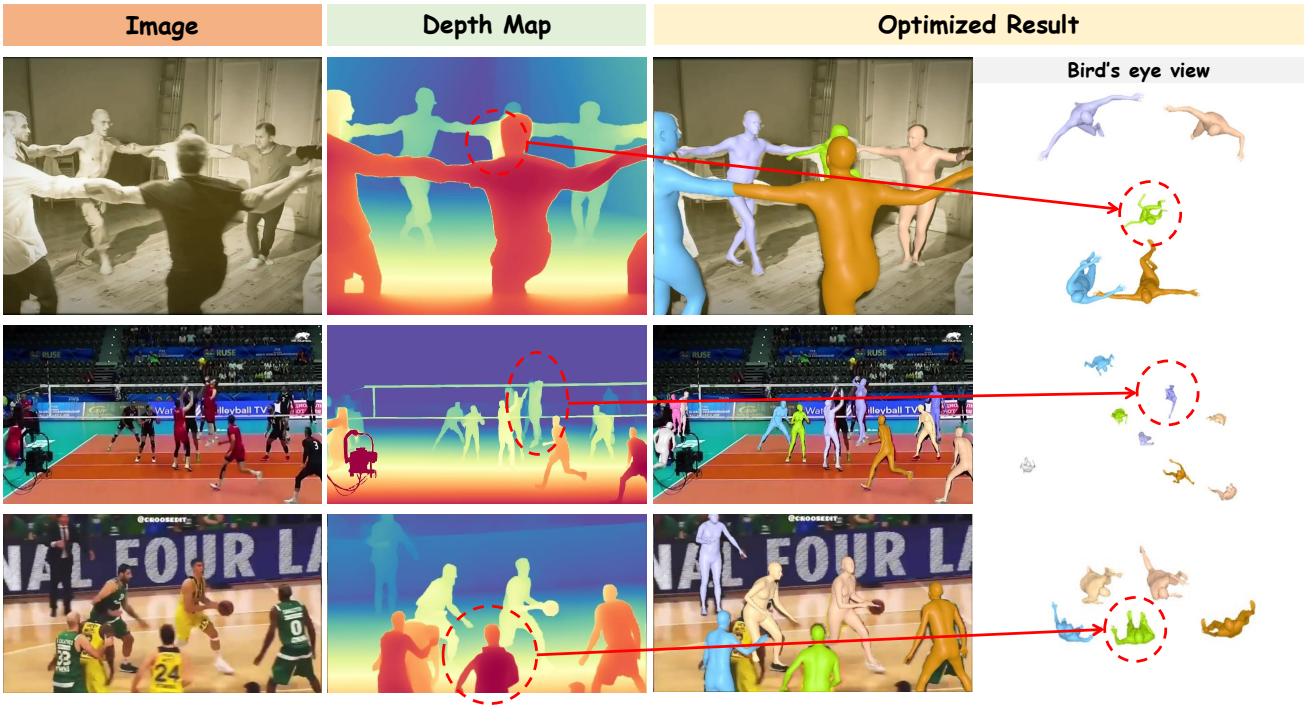

Figure 16. Examples of DTO failure cases.

11. 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. 16 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.

A further limitation lies in our reliance on CDC growth charts and generalized demographic data to construct the statistical height priors. While a universal Gaussian prior $\mathcal{N}(\mu, \sigma^2)$ serves as a strong and effective anchor for recovering metric scale, it inherently smooths over natural demographic variances (e.g., regional or ethnic height differences). Consequently, the global scale of the reconstructed scene may exhibit slight biases depending on the specific demographic makeup of the subjects. Future work could refine this by incorporating localized or demographic-specific height distributions when metadata is available, further tightening the metric accuracy of the optimization.

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

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