

MetricHMSR: Metric Human Mesh and Scene Recovery from Monocular Images

Supplementary Material

6. More Implementation Details

6.1. Camera, Image and Metrics

In human mesh recovery, full perspective projection [2, 17, 28, 53] has gained increasing attention. The development of monocular metric depth estimation [3, 15, 37, 38, 57, 59, 65] and single-image intrinsic parameter estimation [22, 49] has made it possible to recover metric human mesh from a single image. The intrinsic parameters of the camera and the bounding boxes (bbox) are crucial to perceiving human metric information and the 3D position.

Fig. 9 shows the impact of the camera’s intrinsic parameters on the metric depth of the human. Under the same resolution, if two cameras ($f_2 = 2f_1$) produce the same image size for the same person, their metric depths are different ($d_2 = 2d_1$). If the focal length can be provided as a known parameter to the network, it becomes possible for the network to infer the metric depth of human body.

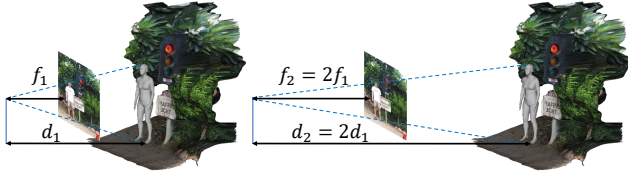


Figure 9. Illustration of the human body’s metric distance under the same 2D projection with different focal lengths. Focal: f , distance: d .

As shown in Fig. 10, the distance between the bounding box and the principal point (c_x, c_y) is directly related to the 3D position of the human body in the camera space.

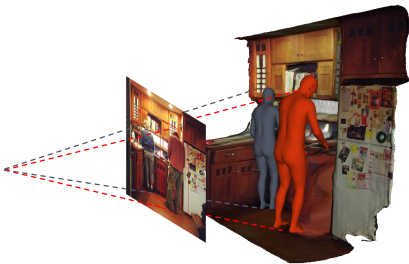


Figure 10. Illustration of the human body’s metric position with different bounding box. The position of the bounding box directly influences the 3D location of the human body.

6.2. Camera Ray Representation

CLIFF [28] was the first to recognize the global information contained in the bounding box, and it incorporated bounding-box as additional input to the network. However, this approach neglects the role of camera intrinsics. To address this limitation, we draw inspiration from NeRF [32] and introduce the bounding ray map representation method, as shown in Fig. 3.

6.3. Losses of Metric Human Mesh Recovery

In Eq. (5), each loss term is computed as

$$\begin{aligned}\mathcal{L}_{J_{2D}} &= ||\hat{J}_{2D} - J_{2D}||_2, \\ \mathcal{L}_{J_{3D}} &= ||\hat{J}_{3D} - J_{3D}||_2, \\ \mathcal{L}_{V_{3D}} &= ||\hat{V}_{3D} - V_{3D}||_2, \\ \mathcal{L}_\theta &= ||\hat{\theta} - \theta||_2, \\ \mathcal{L}_\beta &= ||\hat{\beta} - \beta||_2, \\ \mathcal{L}_h &= ||\hat{h} - h||_2,\end{aligned}\tag{8}$$

where J_{2D} , J_{3D} , V_{3D} , θ , β , and h represent the ground truth of the 2D keypoints, the 3D keypoints, the vertices of the SMPL model, the pose parameters of the SMPL and the shape parameters of the SMPL, respectively. \hat{J}_{2D} , \hat{J}_{3D} , \hat{V}_{3D} , $\hat{\theta}$, $\hat{\beta}$, and \hat{h} represent the corresponding network predictions of the 2D keypoints, the 3D keypoints, the vertices of the SMPL model, the pose parameters of the SMPL and the shape parameters of the SMPL, respectively.

6.4. HumanMoE

On one hand, HumanMoE condenses all 3D attributes into different experts within a single module. On the other hand, it assigns features from different hierarchies and dimensions to specialized experts.

At the patch level, the Patch MoE routes patches of different body parts to distinct experts. As shown in Fig. 5, this leads to semantic decoupling and feature-level disentanglement, enabling specialized processing and improving performance.

At the global image level, the Global MoE routes the aggregated image feature to specialized experts via a single query token. Complementing the Patch MoE, it operates holistically to capture global context and enable consistent reasoning over metric properties for final prediction. Fig. 11 and Fig. 12 illustrate its role.

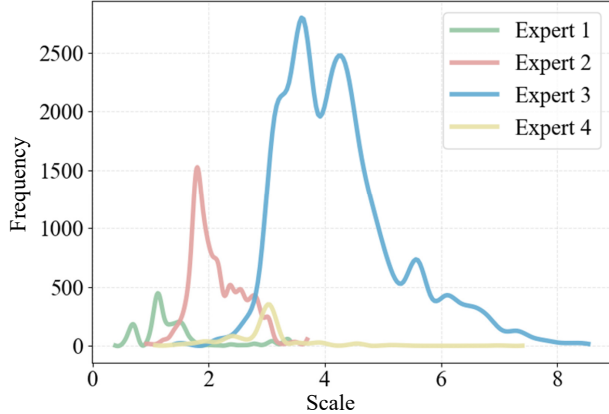


Figure 11. Expert allocation of the global MoE across different bbox sizes. Scale denotes the factor by which the cropped images are resized to a fixed size.

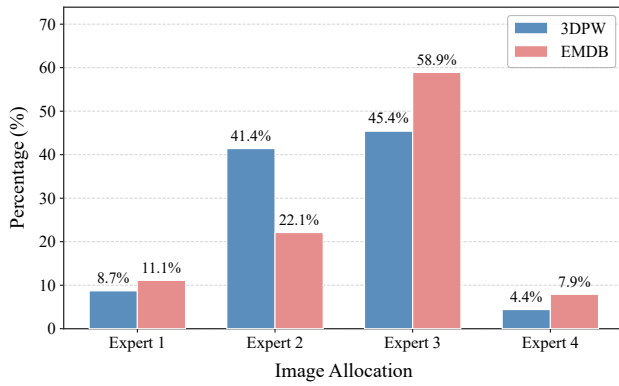


Figure 12. Expert allocation of the global MoE on different datasets.

We tested on the fixed-intrinsics dataset 3DPW. The cropped images are scaled to a fixed size of 256×256 to feed to our network. This scale reflects the size of the bounding box (bbox). Fig. 12 shows how the first layer of the global MoE allocates images in accordance with this scale. It can be observed that images with larger bounding boxes (i.e., closer people) tend to be assigned to Expert 3, while those with smaller bounding boxes (i.e., farther people) are more often assigned to Expert 1.

Fig. 12 shows the assignment of images from the 3DPW and EMDB datasets to experts in the final layer of the MoE. Due to differences in human pose, scene, and other factors across datasets, the allocation to experts also differs.

6.5. Human-Guided Metric Depth Refinement

We refine a predicted dense depth x_{in} into metric-consistent depth by learning per-pixel affine parameters (s, b) so that $x_{out}(u, v) = s(u, v) x_{in}(u, v) + b(u, v)$. Absolute-scale constraints (“anchors”) come from human geometry: at infer-

ence we project an HMR mesh and keep only reliable visible vertices. We then exclude unstable anatomical regions using a precomputed mask, form a dense anchor map by taking per-pixel nearest- Z , and apply a light depth-gradient filter to suppress boundary/occlusion outliers.

The predictor is a UNet encoder-decoder with a ViT bottleneck. The encoder stacks Conv-GroupNorm-GELU blocks with downsampling; the deepest feature is patch-embedded and fed to a Transformer encoder (multi-head self-attention + MLP, norm-first), then reshaped and fused in the decoder via skip connections. Inputs are concatenated maps: the working-domain depth x_{proc} (depth or disparity), the anchor map, the anchor mask, and optional RGB. The head outputs two channels \tilde{s}, \tilde{b} , constrained as $s = 1 + \alpha_s \tanh(\tilde{s})$ (small scale around 1) and $b = \alpha_b \text{median}(|x_{proc}|) \tanh(\tilde{b})$ (scene-adaptive bias). To stabilize training, x_{proc} and anchors are normalized per batch by masked mean-absolute values and clipped.

We use a depth reconstruction loss on valid GT pixels, an anchor consistency loss where anchors exist, and two regularizers on s, b : total variation (spatial smoothness, optionally edge-aware) and spatial variance (keeps fields near global). The total loss is a weighted sum of these terms. Operating in disparity is supported (we convert back to depth for output); tight constraints on s and regularization on s, b preserve near-planarity and prevent over-correction. For training, we split a human-centric RGB-D dataset (PROX) into train/test with a 9:1 ratio.

We would like to clarify that our design is motivated by the observation that human reconstruction from MetricHMR is relatively reliable in metric scale and can therefore serve as a geometric anchor for refining monocular depth estimation. Specifically, we use the reconstructed metric human mesh to guide the refinement of MapAnything depth predictions. Unlike prior approaches that apply a simple global scaling factor, we introduce a per-pixel refinement module to correct spatially varying depth errors. This design accounts for the non-uniform inaccuracies commonly observed in monocular depth estimation.

As shown in Tab. 5, this human-guided refinement improves the metric accuracy of scene depth on the PROX dataset. The current accuracy of monocular depth estimation methods, such as MapAnything, remains insufficient to be reliable. For example, in Fig. 6, the estimated subject height is 1.59 m while the ground-truth height is 1.72 m. Due to this limitation, we use the predicted depth only as an initialization for scene reconstruction rather than as a constraint to further refine the human pose.

7. Additional Results

7.1. SynFocal dataset.

Across images with varying camera parameters, MetricHMSR consistently achieves accurate metric distance perception. Recognizing that existing datasets predominantly feature fixed focal lengths, we construct SynFocal, a synthetic dataset of human images that vary in focal length, to further validate the performance of MetricHMSR.

Our scenes are constructed using 3D-FRONT [14], a synthetic dataset characterized by extensive layout variations and high-fidelity furniture models. Human models are sourced from THuman [60], which comprises a large collection of high-quality 3D human models with corresponding SMPL-X [36] annotations. We used Blender [5] to arrange the human and scene in a plausible configuration and rendered images at different focal lengths. Fig. 13 shows some examples of SynFocal.

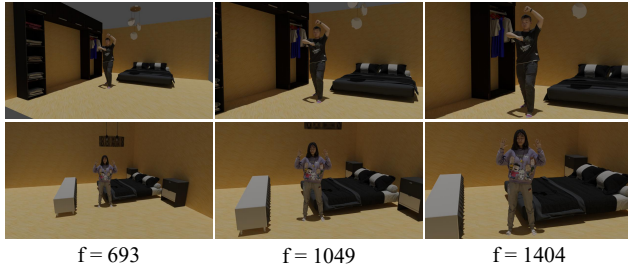


Figure 13. Examples of SynFocal. Each column of images was rendered using a distinct focal length (pixels).

Note that, to the best of our knowledge, there were no human datasets with varying focal lengths before our work. Meanwhile, our concurrent work Bedlam2.0 [46] provides richer camera motion and focal length diversity, which can serve as a more comprehensive alternative to SynFocal.

7.2. Human mesh recovery under varying intrinsics

For quantitative comparisons, we evaluate performance using two metrics: Root Distance Error (RDE) and Root Position Standard Deviation (RPSD). RDE measures accuracy in the camera coordinate system as the Euclidean distance between predicted and ground-truth root positions, normalized by the distance from the ground-truth root to the origin. RPSD measures robustness as the standard deviation of the Euclidean distances from predicted root joints to their centroid. We conduct a systematic comparison between MetricHMSR and Human3R [8] under varying camera intrinsics. Quantitative results are reported in Tab. 7. The Results demonstrate that MetricHMSR exhibits stronger robustness to focal length variations than Human3R.

Method	RDE	RPSD
Human3R [8]	0.14	21.6
MetricHMSR	0.10	1.1

Table 7. Quantitative comparisons of 3D position estimation under different intrinsics. RDE is in % and RPSD is in cm.



Figure 14. Metric human and scene reconstruction results of in-the-wild data from COCO.

7.3. Pseudo-GT Depth for 2D Image Datasets

Based on our Human-Guided Metric Depth Refinement module, we introduce in the main text a new pseudo-GT dataset that supplements background depth information for the commonly used 2D image datasets AI Challenger, COCO, and MPII. Fig. 14 illustrates some examples on in-the-wild data from the COCO dataset. We use Human-FoV [35] to estimate the intrinsics of the image, then reconstruct the metric human mesh along with the scene.

More broadly, our method reconstructs metrically accurate humans and scenes from a single image without requiring video, making it possible to leverage the vast amount of image data available across the internet.

Method	H-MAE	H-MAPE	PVE-T	H Var.	PVE-T Var.
PromptHMR	76.0	4.3	32.4	11.2	67.7
MetricHMSR	70.1	3.9	29.9	4.1	27.9

Table 8. Metrics on 3DPW. Var. denotes variance.

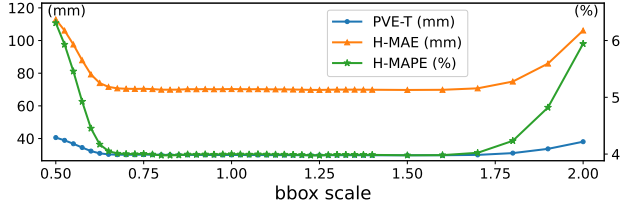


Figure 15. Impact of bounding box scale variation on shape and height accuracy. H: height, T: T-pose, MAE: Mean Absolute Error, MAPE: Mean Absolute Percentage Error.

7.4. Metric Comparison

The metric properties mainly concern metric 3D position and metric body shape. Tab. 8 reports the error and stability of the shape, showing that our method outperforms SOTA.

7.5. Impact of Bbox Scale

Fig. 15 plots the variation of the metric against the bbox, where a scale of 1.0 denotes the standard bbox. It is evident that the proposed method demonstrates strong robustness to variations in the bbox.

7.6. Trajectory Comparison

In the quantitative comparisons presented in the main text, our method significantly outperforms existing online approaches in terms of trajectory accuracy, and achieves competitive performance with video-based methods while operating on frame-by-frame inputs rather than video sequences.

Fig. 16 shows qualitative comparisons with existing online method on RICH. Our method demonstrates stronger scale awareness and reduced positional drift.

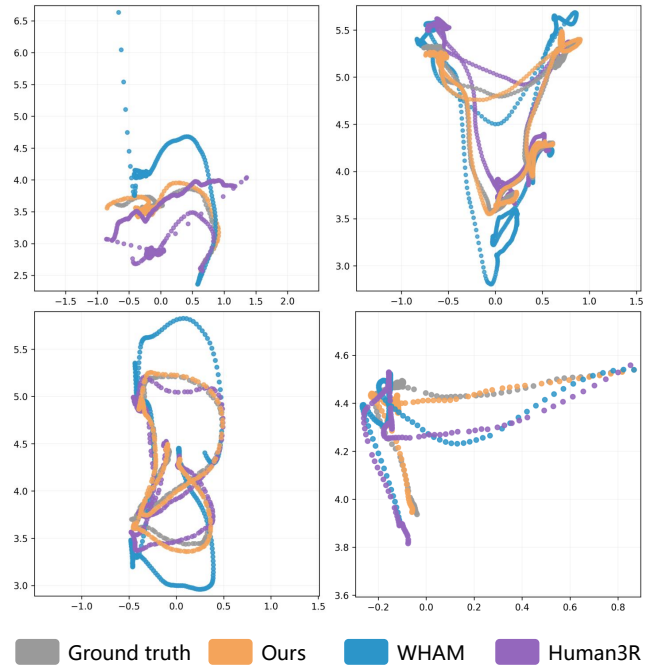


Figure 16. Qualitative comparison of global motion trajectories.