Supplementary Material

Loss Weight Selection

The loss weights λ_1, λ_2, and λ_3 were selected to balance the contributions of the segmentation, DensePose, and binary mask losses, respectively. To ensure all loss terms operate on a comparable numerical scale, we first computed the typical magnitude of each loss on a held-out validation set and then scaled the corresponding λ accordingly. The final values used in our experiments were $\lambda_1 = 0.01$, $\lambda_2 = 1$, and $\lambda_3 = 1$.

Runtime Comparison

To contextualize the efficiency of our approach, we report end-to-end runtimes on a representative 4-view input of a 3-person scene (EgoExo4D-style setup). All measurements were performed on a single NVIDIA V100 GPU.
HSfM uses a multi-stage optimization pipeline with external tools, each adding to the total runtime. In contrast, HAMSt3R is a unified, feed-forward architecture with optional SMPL fitting used only for evaluation.

Despite being a unified feed-forward model, HAMSt3R remains highly efficient compared to modular pipelines such as HSfM. SMPL fitting is performed only as a post-processing step for evaluation.

Table 7. **Runtime comparison** for HAMSt3R and HSfM in a 4-view, 3-person scene. HAMSt3R is significantly faster despite producing comparable or stronger results.

Method	Total Runtime (4-view)
HAMSt3R (Ours)	
Reconstruction + Segmentation + DensePose	∼14s
SMPL Fitting (post-process)	∼6s
Total	**∼32s**
HSfM [41]	
2D Pose Initialization	∼1s
Segmentation (SAM [47])	∼2s
3D Reconstruction (DUSt3R [64])	∼18s
Stage 1 (Translation & Scale Only)	∼25s
Stage 2 (Add Global Orientation & Align DUSt3R)	∼48s
Stage 3 (Add Local Body Pose)	∼24s
Total	**∼118s**

Monocular Prediction

HAMSt3R can also be run on a single image, by simply feeding the same image twice to the network. We show some qualitative results on in-the-wild images in Figure 6.

Failure Cases in SMPL Fitting

While HAMSt3R is generally robust, we observe occasional failures in SMPL fitting, especially when subjects are far from the camera. In such cases, the 3D points are sparse and noisy, leading to unstable optimization (Figure 7).

We hypothesize that increasing reconstruction resolution or incorporating additional priors could improve robustness in these edge cases. These examples highlight the challenges of downstream mesh fitting in low-density areas, even when the upstream reconstruction is geometrically correct.

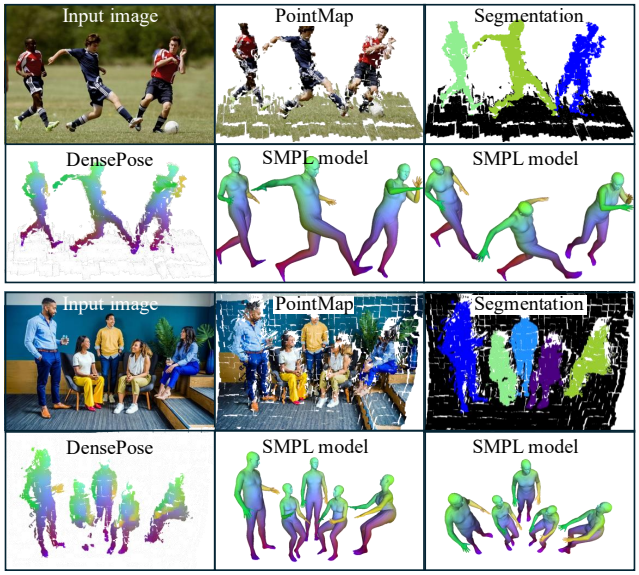

Figure 6. **Qualitative results of monocular prediction** on in-the-wild images taken from Pexels [2]. Each 2-row group shows: (top row, left to right) input image, high-confidence reconstructed point cloud overlaid with color (PointMap), and segmentation results; (bottom row) DensePose predictions and two different views of the fitted SMPL models. The figure illustrates that our method can produce coherent reconstructions and mesh predictions from a single image.

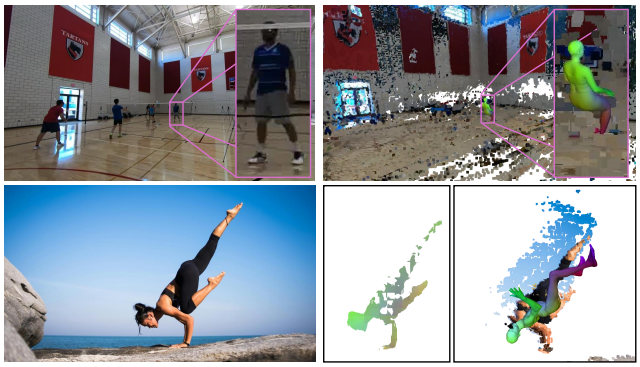

Figure 7. **Examples of failure cases.** *Top row*: reconstruction failures on a scene from EgoHumans. Due to the large scene scale and wide camera angle (left), human subjects appear very small after downscaling, leading to noisy point clouds (right) and incorrect orientation in the SMPL mesh (inset). *Bottom row*: failures caused by extreme body poses. DensePose predictions (middle) can break down under uncommon configurations (left), resulting in erroneous SMPL fits (right).