# WILDHSR: METRIC FEED-FORWARD 4D PEOPLE-SCENE RECONSTRUCTION FROM A 3D FOUNDATION MODEL

**Anonymous authors**
Paper under double-blind review

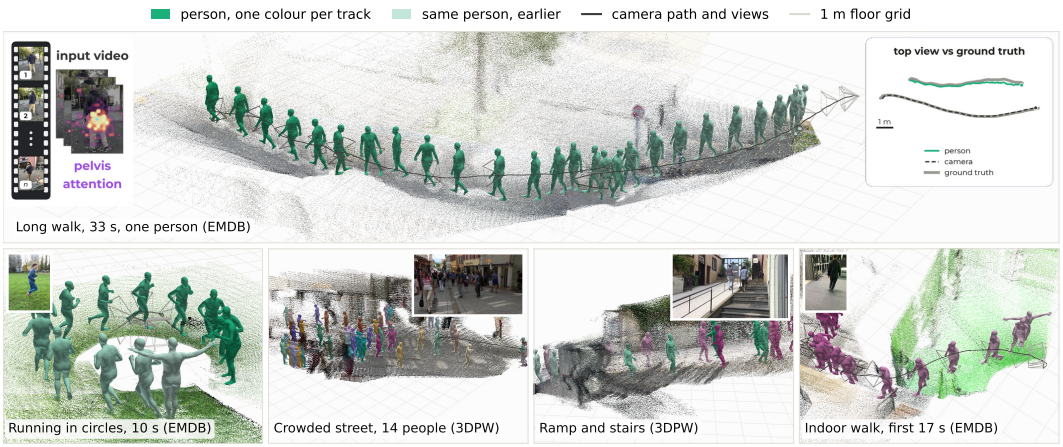



Figure 1: **Metric people, scene and camera from one moving camera.** Each panel overlays persistent people, the person-free scene and the camera path on a 1 m grid. The large EMDB example spans 33 s; its insets show pelvis-token attention and a top view against ground truth after rigid rotation-and-translation alignment, preserving the predicted scale. The remaining clips cover cyclic motion, a 14-person street, stairs and an indoor walk. Inference receives a single monocular video.

## ABSTRACT

3D foundation models recover video cameras and geometry in one forward pass, but some of the strongest are up to scale. Joint people-scene reconstruction then requires two missing outputs: *metric scale* and persistent *person identity*. We ask whether one up-to-scale foundation representation can support both through lightweight adaptation. Exact metric labels are scarce, but unlabeled in-the-wild video is abundant. We use people in curated web video to initialise the solution: a posed metric body and 2D keypoints give an approximate, closed-form scale pseudo-label. These pseudo-labels pretrain a Scale Readout, which is then fine-tuned together with a lightweight adapter using exact metric supervision from standard real-video training splits. At inference the Scale Readout predicts metric scale from foundation-model tokens, without the ruler or its teachers. For *person identity*, we probe the pretrained foundation model alone and find evidence that its intermediate query-key features encode person correspondence across frames. In most evaluated moving-person clips, a mid-layer token prefers that person over the vacated location and other people. A tiny projection reads this correspondence; together with metric pelvis motion and proposal confidence, it drives dustbin-aware Sinkhorn association of per-frame bodies. *WildHSR* combines both readouts to reconstruct metric cameras, scene and people from monocular video. On EMDB-2, *WildHSR* is the first feed-forward method in the published comparison to beat the best optimization-based WA-MPJPE and RTE while leading feed-forward methods on all three world-frame metrics. On RICH, it leads feed-forward people-and-scene methods on WA-MPJPE and W-MPJPE. The complete pipeline runs at 10.1 fps on one GPU.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

# 1 INTRODUCTION

Reconstructing people together with their surroundings from monocular video requires bodies, cameras and scene geometry in one metric world. 3D foundation models now recover strong relative geometry and camera motion in a single forward pass (Wang et al., 2024a; 2025b;a; 2026), including video whose content moves (Zhang et al., 2025). Some learn metric output from metric supervision (Wang et al., 2025b; Keetha et al., 2026; Yang et al., 2026); some of the strongest reported camera estimates instead come from models whose output is up to scale (Wang et al., 2025a; 2026), and it is such a representation we study. Joint people-scene reconstruction then lacks two quantities. First, monocular geometry is ambiguous up to a global *metric scale* (Eigen et al., 2014), and an up-to-scale model normalises each window to an arbitrary unit. Second, its geometric output has no persistent *person identity*: it does not say which body in one frame is the same person later. Over time, independently scaled windows distort trajectories, while identity switches splice different people into one path.

The prevailing answer is to obtain both outside the up-to-scale representation: metric scale from a metric-native backbone (Chen et al., 2026; Wang et al., 2025b), from metric depth priors (Wang et al., 2024b), or from synthetic metric supervision and a depth teacher (Li et al., 2026; Shi et al., 2026), and person identity from external detectors and trackers (Ye et al., 2023; Shin et al., 2024; Shen et al., 2024). These choices introduce separate models, supervision sources or sequence-processing stages. We ask whether one up-to-scale foundation representation can support both capabilities without an external scale network or separately trained tracker. Scale requires supervision. Person correspondence presents a different opportunity: 3D foundation models learn matching and geometric consistency across views, so their intermediate representations may retain correspondence even on moving people. Large vision models often encode structure beyond their explicit outputs: object segmentation emerges in the attention of self-supervised transformers (Caron et al., 2021), 3D structure can be probed out of image foundation models (El Banani et al., 2024), and the attention of a pairwise geometry model separates moving objects from the static scene with no training at all (Chen et al., 2025). We therefore transfer automatic human-derived scale supervision into its scene tokens and test whether person correspondence can be read from its pretrained representation (Alain & Bengio, 2017).

**Reading person correspondence from the representation.** Reconstructing a body in each frame does not establish which bodies belong to the same person over time. Rather than importing a separate visual tracker, we ask whether the 3D backbone itself carries that correspondence. It learns geometric matching across views, yet moving people are excluded from its matching supervision (Wang et al., 2026). We probe the unmodified backbone before training an identity readout and find a correspondence signal on people in its intermediate query-key features. Figure 2 illustrates the finding: patches on each person in one frame preferentially match patches on that person later, forming same-person blocks. A retrieval probe validates the layer choice: intermediate query-key features retrieve the person even after motion, while early and late features are nearer chance. This identifies a useful layer for reading identity. A small projection extracts person features from the intermediate representation; metric pelvis motion and proposal confidence then help resolve ambiguous links through dustbin-aware analytic assignment. Thus the same 3D representation that supports people and scene reconstruction also supplies the correspondence cue for persistent tracks, without a separately trained visual tracker.

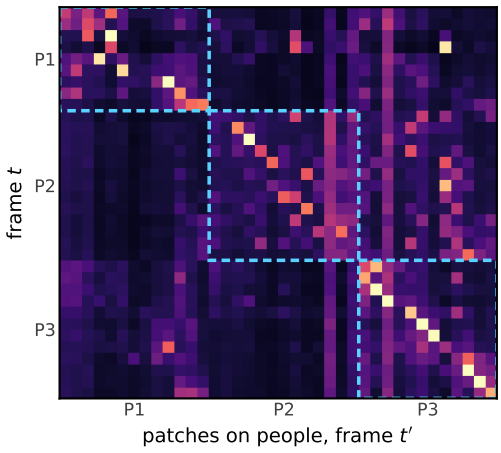



Figure 2: **Person-patch matching at layer 13.** Rows are patches on three people at $t$; columns are patches on the same people at $t'$. Color shows pre-rotation query-key probe scores, normalized only over these candidates per head and then averaged, not scene-wide attention. Dashed blocks mark same-person pairs; 82% of top matches have the right identity here.

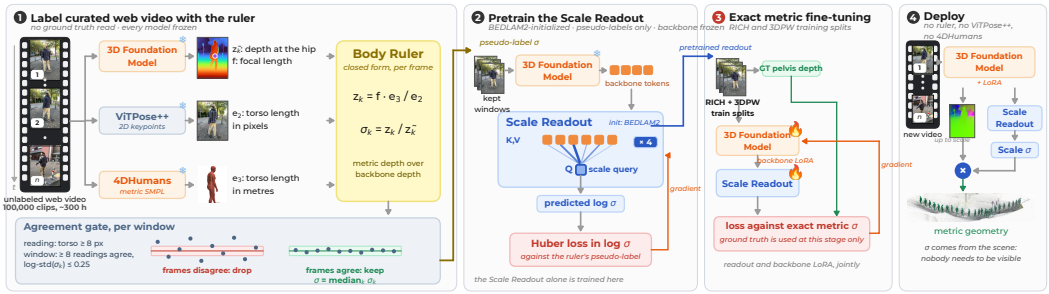



Figure 3: **Training and deploying the Scale Readout.** (1) A body ruler combines observed 2D torso extent $e_2$, the posed body's metric in-plane extent $e_3$, backbone depth $\tilde{z}_k$ and focal length $f$ to estimate $\sigma_k = (fe_3/e_2)/\tilde{z}_k$; an agreement gate yields window pseudo-labels. (2) These labels pretrain the readout on web clips. (3) Exact RICH and 3DPW targets fine-tune the readout and backbone LoRA. (4) Deployment predicts $\sigma$ from backbone tokens alone, without the ruler or teachers.

**Teaching metric scale to scene tokens.** Exact metric labels for ordinary real video are scarce. We instead use people as offline rulers: a posed body supplies a metric torso extent in the image plane (Loper et al., 2015; Pavlakos et al., 2019), which we compare with 2D keypoints to obtain an approximate, closed-form scale pseudo-label (Fig. 3). Repeated readings expose inconsistent labels, enabling pretraining on curated in-the-wild (ITW) video before exact metric fine-tuning with low-rank adaptation (LoRA) on standard real-video training splits. The resulting Scale Readout predicts from scene-level foundation-model tokens rather than repeating the body measurement at inference. This transfers human-derived supervision into a scale estimate that is independent of a visible person at deployment.

*WildHSR* is the system these two mechanisms make possible: metric cameras, scene and people from monocular video, with no external scale network or separately trained external tracker (Fig. 4). Our contributions are:

- ***WildHSR***, a system that jointly reconstructs people, cameras and scene geometry in one metric world from monocular video while maintaining person tracks.
- **Emergent person correspondence in a 3D foundation model**: a controlled probe locates motion-robust identity cues in intermediate query-key features; a lightweight projection reads them for association without an external tracker (§3.2).
- **Human-derived scale supervision for scene tokens**: closed-form pseudo-labels from people in unlabeled video initialise a Scale Readout before exact metric adaptation (§3.3, Table 3).
- **State-of-the-art global motion**: on EMDB-2, the first feed-forward method in our comparison to beat the best optimization method on WA-MPJPE and RTE; on RICH, the best feed-forward people-and-scene method on WA-MPJPE and W-MPJPE.

## 2 Related Work

**3D foundation models, and what they encode.** Feed-forward geometry models recover cameras and dense structure from unposed images in one pass (Wang et al., 2024a; 2025a), online with a persistent state (Wang et al., 2025b) and in the presence of motion (Zhang et al., 2025). Those that output metres learn them from metric supervision (Wang et al., 2025b; Keetha et al., 2026; Yang et al., 2026); some of the strongest reported camera estimates come from a model that is up to scale and re-normalised per inference (Wang et al., 2025a; 2026). We build on that representation and learn its metric conversion. What such models encode beyond their outputs is less studied. Image foundation models carry 3D structure that probes can read (El Banani et al., 2024), segmentation emerges in self-supervised attention (Caron et al., 2021), DUSt3R's attention separates moving from static content without training (Chen et al., 2025), and the authors of VGGT-Ω report that clustering its intermediate tokens isolates a moving dancer, and that auxiliary quantities, metric scale among them, can be decoded from its register tokens in preliminary experiments (Wang et al., 2026). We

test two readouts: *metric scale* from in-the-wild pseudo-labels followed by exact adaptation, and *person identity* from cross-frame correspondence in intermediate features.

**Supervision for metric scale.** Systems that reconstruct people and scenes in metres obtain metric scale from a metric-native backbone (Chen et al., 2026), from metric depth priors (Wang et al., 2024b), from synthetic metric supervision with an expert depth teacher (Li et al., 2026), or by training the body's scale prior into point-map prediction (Shi et al., 2026). Anthropometric size is an older cue. People as Scene Probes (Wang et al., 2020) reads depth, occlusion and lighting from passing pedestrians, SLAHMR and PACE use body priors inside their objectives (Ye et al., 2023; Kocabas et al., 2024), HAMSt3R (Rojas et al., 2025) distils a mesh recovery encoder into a stereo network, HSfM (Müller et al., 2025) recovers approximate metric scale by per-scene optimization and, closest in mechanism, HAC (Yang et al., 2025) calibrates a SLAM reconstruction against the metric depth of contact joints from a mesh recovery model. In these uses, the body measurement remains part of scale recovery at inference. We instead use the reading offline to pretrain on curated unlabeled video, in the tradition of self-training (Lee, 2013; Xie et al., 2020) and consistency supervision (Godard et al., 2019), then refine the predictor with exact metric labels from real-video training splits. UniSH also trains on unlabeled in-the-wild video, but uses an external metric teacher; our pseudo-label is produced by the geometric body ruler.

**Humans and scenes from video.** Global human mesh recovery places SMPL (Loper et al., 2015) bodies in world coordinates, by optimization over SLAM and motion priors (Ye et al., 2023; Kocabas et al., 2024) or by regressing world trajectories (Yuan et al., 2022; Shin et al., 2024; Wang et al., 2024b; Shen et al., 2024; Wang et al., 2025c). Closest to us, Human3R (Chen et al., 2026), UniSH (Li et al., 2026), SHOW (Shi et al., 2026) and GUSH3R (Abe et al., 2026) attach human decoders to geometry foundation models and reconstruct people and scene in one feed-forward pass, the last as Gaussians; MetricHMSR (Song et al., 2025) does so metrically from one image, and JOSH3R (Liu et al., 2026) is trained from the pseudo-labels of a per-sequence optimization. These systems differ in where the human branch reads the backbone. Human3R decodes SMPL-X parameters at a detected head cell, which becomes ambiguous when two people share that token. We read each person at the pelvis, a geometric anchor for body placement (§3.2).

# 3 METHOD

## 3.1 PROBLEM FORMULATION AND SYSTEM OVERVIEW

A window of video passes once through a 3D foundation model (VGGT-Ω (Wang et al., 2026), the successor of VGGT (Wang et al., 2025a)), adapted with a small LoRA for metric transfer, which predicts per-frame cameras with focal length $f$ and depth $\tilde{z}$, from which point maps are obtained by unprojection, together with the tokens used to decode these outputs (Fig. 4, top). All of it is correct up to one unknown scale. We write $\sigma = z_{\text{metric}}/z_{\text{backbone}}$ for the metres-per-unit conversion of that inference. The backbone re-normalises every pass, so $\sigma$ must be predicted per window rather than pooled over a sequence. Cameras and scene are in backbone units, whereas SMPL-X dimensions are metric; $\sigma$ puts both in one world.

Two pretrained networks supply scene and human tokens, while our Scale Readout, Pelvis Readout, cross-attention fusion and identity projection expose the quantities needed for joint reconstruction. Their released weights remain frozen; small LoRA adapters support metric and body adaptation. Appendix B gives the exact modules, parameter counts and training configuration. We next describe how the system reads and associates **people**, learns **scale**, and composes one metric **world**.

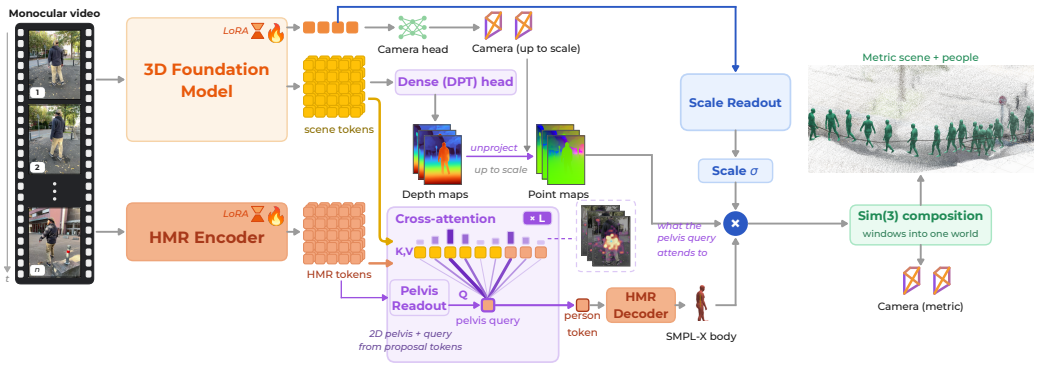



Figure 4: Overview of ***WildHSR.*** A 3D foundation model produces up-to-scale cameras, geometry and scene tokens. A mesh branch proposes people; a Pelvis Readout forms a body-localized query, and cross-attention decodes a metric SMPL-X body. The Scale Readout supplies $\sigma$; analytic association and Sim(3) composition link people and windows in one metric world.

### 3.2 Person reconstruction and temporal association

**Per-frame proposals and bodies.** Multi-HMR (Baradel et al., 2024), adapted with a 0.75M LoRA, scores person-centre patches and proposes initial SMPL-X bodies per frame; VGGT-$\Omega$ independently supplies scene tokens for the video window. For proposal $n$ at time $t$, a lightweight Pelvis Readout takes its HMR tokens and produces a 2D pelvis location and person query. During training, the location is supervised by the pelvis joint of the ground-truth SMPL-X body projected into the image; the target does not come from the model's own final prediction. We then fuse the readout query with the two token streams:

$$\begin{aligned}(u_{t,n}, q^{\mathrm{pel}}_{t,n}) &= R_{\mathrm{pel}}(H_{t,n}),\\ h_{t,n} &= \mathrm{CrossAttn}\Big(q^{\mathrm{pel}}_{t,n}, [S_{1:T}; H_t]\Big), \quad B_{t,n} = D_{\mathrm{HMR}}(h_{t,n}).\end{aligned} \tag{1}$$

Here $H_{t,n}$ denotes the HMR tokens of proposal $n$, $u_{t,n}$ is the readout's image-space pelvis location, $q^{\mathrm{pel}}_{t,n}$ its query, $S_{1:T}$ the VGGT scene tokens, and $H_t$ the frame's HMR tokens; the latter two supply keys and values. At inference, the Pelvis Readout needs no ground-truth body or external keypoint model. No external detector runs in the deployed path.

**The pelvis anchor.** Placement unprojects the Pelvis Readout's pixel $u_{t,n}$ using the median backbone depth $\tilde{z}_{t,n}$ over the body's projected torso. With backbone-unit camera centre $C_t$, camera-to-window rotation $R_t$, intrinsics $K_t$ and $\bar{u}_{t,n} = [u_{t,n}^{\mathsf{T}}, 1]^{\mathsf{T}}$, the metric pelvis in window $w$ is

$$p^{(w)}_{t,n} = \sigma\big(C_t + R_t(\tilde{z}_{t,n} K_t^{-1} \bar{u}_{t,n})\big). \tag{2}$$

The readout is supervised at the SMPL-X pelvis joint, so $p^{(w)}_{t,n}$ places the body's root. The torso median avoids a body-ground depth discontinuity at the anchor.

**Identity readout and association.** The probe of Fig. 5 peaks in the layer-13 query-key space. We pool those vectors around each pelvis and map them to a 128-dimensional unit descriptor with a two-layer projection trained by supervised contrastive loss. For track $i$ and proposal $j$, let $m_i, e_j$ be unit descriptors, $\hat{p}_i, p_j$ their predicted and observed metric pelvis positions, $c_j$ proposal confidence and $g_i$ a time-dependent motion gate. Their assignment cost is

$$C_{ij} = \tfrac{1}{2}(1 - \langle m_i, e_j \rangle) + \min\left(\frac{\|p_j - \hat{p}_i\|_2}{g_i}, 1\right) + 0.25(1 - c_j). \tag{3}$$

Pairs beyond $g_i$ are invalid; Appendix I specifies the gate and track state. A dustbin and Sinkhorn optimal transport (Cuturi, 2013) give a soft assignment, which is hardened one-to-one with Hungarian matching. The projection also receives a ground-truth assignment loss through the soft Sinkhorn matrix; Hungarian selection and confidence-gated memory updates remain outside backpropagation. Appendix I gives the full specification. Learned inference is feed-forward within each window; fixed-step association and memory updates require no test-time gradient-based fitting.

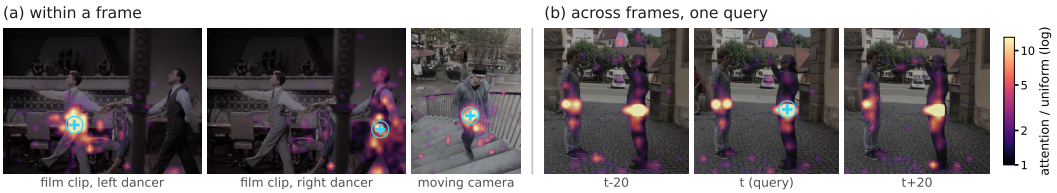



Figure 5: **Pretrained person correspondence.** At layer 13, a VGGT-Ω pelvis token follows its own person across frames: once the person moves, its attention on the new position is $5.0\times$ uniform, against $1.2\times$ on the vacated location. This probe runs the backbone alone; Appendix I gives controls and the complete sample accounting.

### 3.3 Scale Readout and Metric Adaptation

**Ruler pseudo-labels.** For each visible person, ViTPose++-H (Xu et al., 2024) at $256 \times 192$ gives the observed image-plane extent $e_2$ from shoulder midpoint to hip midpoint. 4DHumans (Goel et al., 2023) supplies the posed metric torso; $e_3$ is the in-plane extent between its corresponding midpoints. With the backbone's focal length $f$ and estimated depth $\tilde{z}_k$, a local weak-perspective approximation gives the ruler reading (Fig. 3, stage 1)

$$z_k = f\,\frac{e_3}{e_2}, \qquad \sigma_k = \frac{z_k}{\tilde{z}_k}, \tag{4}$$

without a metric label or sensor for the web clip. The estimate can be biased when the torso endpoints have different depths. We retain confident detections with $e_2 \geq 8\,\text{px}$; a window needs at least eight readings and $\mathrm{std}(\log \sigma_k) \leq 0.25$. Its pseudo-label is the median across eligible frames and people; agreement tests consistency, not absolute correctness. One visible person can provide all eight readings across frames. More people add candidate readings to the same median and may reduce sensitivity to an individual bad pose or occlusion, but crowding and shared body-prior bias mean that more people do not guarantee a better scale estimate.

**Pretraining and adaptation.** These labels pretrain the Scale Readout on 100,000 curated web-video clips (about 300 hours). The clips include publicly accessible video such as YouTube; YOLO person counts guide sampling across crowd sizes. A learned query reads the backbone's camera and register tokens to regress $\log \sigma$ with a log-space Huber loss while the backbone remains fixed (architecture in Appendix B). After synthetic initialization and ruler pretraining, we fine-tune the readout and backbone LoRA with exact metric targets from standard RICH (Huang et al., 2022) and 3DPW (von Marcard et al., 2018) training splits (Chen et al., 2026). Camera calibration and body translation provide reference pelvis depth; no scene point cloud is needed. At inference, the readout needs only backbone tokens, with no ruler, teacher or visible person.

### 3.4 Composing One Metric World

Given $\sigma$, the backbone's camera centres, depth and point maps are multiplied by it, while the metric body is not, and one forward pass per window yields the metric scene, the metric camera path and the people placed in it (Fig. 4). Long video needs one continuous world frame, but every window is an independent inference with its own frame and unit. We use 100-frame windows with stride 50 and require at least eight shared frames.

For consecutive windows $a$ and $b$, the relative rotation $R_{ba}$ is the chordal mean of the framewise rotations between their shared camera orientations. With $R_{ba}$ fixed, we solve

$$(s_{ba}, t_{ba}) = \arg\min_{s,t} \sum_{k \in a \cap b} \left\| C_k^a - (s R_{ba} C_k^b + t) \right\|_2^2, \tag{5}$$

in closed form on their metric camera centres. If fewer than eight correspondences survive or the centred camera trajectory is degenerate, we retain the previous cumulative similarity. All shared frames have equal weight and no outlier trimming is applied. We compose valid relative similarities in temporal order. If $S_w$ is the cumulative scale of window $w$, we divide every cumulative similarity by $g = \exp\big(|\mathcal{W}|^{-1} \sum_w \log S_w\big)$, preserving the sequence's aggregate metric scale rather than letting one window set it. The similarity acts fully on camera centres, point maps and pelvis translations; its rotation also acts on camera and SMPL-X global orientations, while body dimensions remain unscaled and no network or scene parameters are optimized at test time.

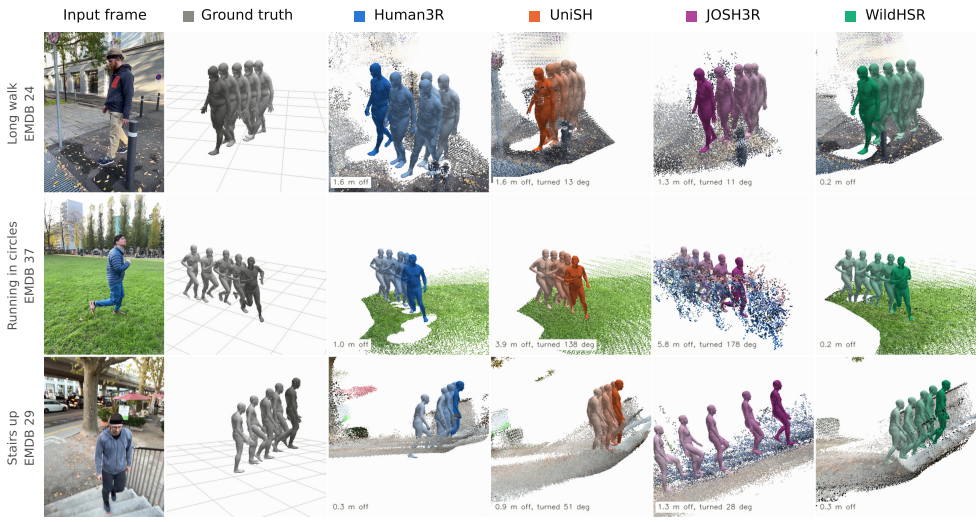



Figure 6: **Scene and people in one metric world** (three EMDB-2 clips), each method at its own metric scale; labels give placement error and, for UniSH and JOSH3R, view turn. *WildHSR* is 0.2 to 0.3 m off, against 0.3 to 1.6 m (Human3R), 0.9 to 3.9 m (UniSH) and 1.3 to 5.8 m (JOSH3R).

### 3.5 Training objectives

The scale pathway uses the log-space Huber objective above; the Pelvis Readout's location is supervised by projected ground-truth SMPL-X pelvis joints, while its query and the fusion receive gradients through the decoded body. No target is imposed on the cross-attention weights themselves. Body losses cover SMPL-X parameters, mesh and reprojection; identity losses use contrastive and soft assignment, while scene contact regularizes placement.

**Feet on the reconstructed ground.** A scene-contact consistency term encourages supporting feet to agree with the local reconstructed surface during training. The body branch predicts its placement directly at inference, without a manual vertical shift (Appendix H).

## 4 Experiments

**Protocol.** We follow the published EMDB-2 (Kaufmann et al., 2023) 25-sequence and RICH (Huang et al., 2022) protocols used by the compared methods (Chen et al., 2026; Li et al., 2026; Shi et al., 2026; Liu et al., 2026; Hu et al., 2025; Sun et al., 2023; Li et al., 2024). WA-MPJPE uses similarity alignment, W-MPJPE aligns the first two frames, and RTE aligns trajectories by rotation and translation only; RTE therefore preserves scale error.

**EMDB-2.** Table 1 compares *WildHSR* with previously reported EMDB-2 results. *Trajectory error is the best published*: RTE 0.9 improves on the prior best (1.3) by 31%. On the joint metrics only JOSH (Liu et al., 2026), which optimizes each sequence jointly over scene and body, is ahead on W-MPJPE (174.7 against 193.6). *WildHSR* has the lowest WA-MPJPE in the full table (66.3). *Among feed-forward methods WildHSR leads all three metrics*, by 38%, 26% and 47% over the best of the others on each. What differs is what each system spends to become metric: metric depth priors, large-scale supervised pretraining or a metric-native backbone, where *WildHSR* learns scale from the people. Fig. 6 shows the reconstructions behind these numbers, and Appendix F the paths themselves (Fig. 9) and nine more clips. The visual comparison is restricted to systems whose neural reconstruction networks jointly predict people and scene: Human3R, UniSH and JOSH3R (Liu et al., 2026); SHOW (Shi et al., 2026) has not released weights and appears in the tables only (alignment and re-run details in Appendix F). Because RTE preserves scale error, this comparison tests the shared metric scene and pelvis-based body placement, not only local pose quality.

**RICH.** *WildHSR* has the best WA-MPJPE of Table 2 and leads the feed-forward people-and-scene methods on W-MPJPE, where only JOSH, an optimization, is ahead. Its median multiplicative scale error is 7.2%. Its per-segment RTE is second among feed-forward people-and-scene systems (2.7),

Table 1: **EMDB-2 global motion.** WA/W: joint error; RTE: trajectory error; **best** 2nd 3rd throughout.

| | method | WA ↓ | W ↓ | RTE ↓ |
|---|---|---|---|---|
| Optimization-based | SLAHMR | 326.9 | 776.1 | 10.2 |
| | COIN | 152.8 | 407.3 | 3.5 |
| | TRAM | 76.4 | 222.4 | 1.4 |
| | PromptHMR-vid | 71.0 | 216.5 | 1.3 |
| | JOSH | 68.9 | **174.7** | 1.3 |
| Feed-forward | WHAM | 135.6 | 354.8 | 6.0 |
| | GVHMR | 111.0 | 276.5 | 2.0 |
| | WATCH | 106.4 | 269.3 | 1.7 |
| | JOSH3R | 220.0 | 661.7 | 13.1 |
| | Human3R | 112.2 | 267.9 | 2.2 |
| | UniSH | 118.5 | 270.1 | 5.8 |
| | SHOW | 109.1 | 262.3 | 2.1 |
| | ***WildHSR*** | **66.3** | 193.6 | **0.9** |
| | Δ vs best | -3.8% | +10.8% | -30.8% |

Table 2: **RICH global motion.** WA/W: global joint error; RTE: per-segment trajectory error.

| method | WA ↓ | W ↓ | RTE ↓ |
|---|---|---|---|
| *Optimization-based* | | | |
| TRAM | 127.8 | 238.0 | 6.0 |
| JOSH | 89.0 | **132.5** | 3.0 |
| *Feed-forward, people and scene* | | | |
| Human3R | 110.0 | 184.9 | 3.3 |
| UniSH | 118.1 | 183.2 | 4.8 |
| SHOW | 107.3 | 172.7 | **2.2** |
| ***WildHSR*** | **73.2** | 163.1 | 2.7 |
| Δ vs best | -17.8% | +23.1% | +22.7% |

Table 3: **Training-data ablation on EMDB-2.** End-to-end joint-position (WA, W) and trajectory (RTE) errors for checkpoints trained with the indicated data sources.

| | training data | | | EMDB-2 | | |
|---|---|---|---|---|---|---|
| checkpoint | BEDLAM2 | curated ITW | RICH+3DPW | WA ↓ | W ↓ | RTE ↓ |
| synthetic base | ✓ | | | 182.2 | 1009.2 | 11.3 |
| ITW pretrained | ✓ | ✓ | | 136.7 | 418.6 | 4.6 |
| no ITW pretraining | ✓ | | ✓ | 116.3 | 255.9 | 3.7 |
| **metric fine-tuned** | ✓ | ✓ | ✓ | **66.3** | **193.6** | **0.9** |

Table 4: **Held-out 3DPW body reconstruction.** 14-joint test errors in mm; published baselines from SHOW (Shi et al., 2026).

| method | PA-MPJPE↓ | MPJPE↓ | PVE↓ |
|---|---|---|---|
| Human3R | 44.1 | 71.2 | 84.9 |
| UniSH | 48.8 | 75.6 | 88.8 |
| SHOW | 41.0 | 67.7 | 78.7 |
| **WildHSR** | **38.7** | **65.2** | **74.4** |

**Local body accuracy.** Table 4 evaluates held-out 3DPW. PA-MPJPE removes per-pose similarity; MPJPE and PVE measure camera-frame joints and mesh. *WildHSR* improves on SHOW, the strongest baseline, by 2.3, 2.5 and 4.3 mm, respectively. These metrics do not assess world placement. Together with Tables 1 and 2, they show that *WildHSR* gains global accuracy without sacrificing local body fidelity.

behind SHOW at 2.2; subjects are near-stationary and RTE divides by a median displacement of 0.52 m. Our model is fine-tuned with exact labels from RICH-train and 3DPW-train; the RICH test partition is held out. Table 2 uses the same test split and evaluation protocol as the baselines, not necessarily the same training data. With camera motion largely removed, the joint results are consistent with the Scale Readout and pelvis-to-scene fusion placing bodies in a shared metric frame.

**Pretrained person correspondence.** Figures 2 and 5 show intermediate features matching a moving person across frames. Tested on 1,500 clips, same-person retrieval reaches 80 to 82% in layers 11 to 15 against 46% chance. This retrieval accuracy is distinct from the conditional match weights in Fig. 2 and the attention lift in Fig. 5. The identity projection reads these mid-depth features and combines them with metric pelvis motion for cross-frame assignment (Appendix I).

**Crowds in the wild.** Fig. 7 shows a held-out 3DPW street scene filmed with a moving hand-held camera, with passers-by entering, leaving and overlapping. *WildHSR* keeps 14 tracks over 10 s and places 10 to 11 people per frame in one metric scene, though only two are annotated; for them, hip-depth error is 0.29 m versus 1.20 m for Human3R, with similar local pose.

**Long video and runtime.** Over 1000 frames (33 s), *WildHSR*'s camera-to-person distance error is 5.4 to 6.0% (9.4 to 12.9% for Human3R); without alignment, its median range error is 14.1 cm and its proxemic zone is correct on 100% of frames. The complete pipeline runs at 10.1 fps on one RTX PRO 6000 Blackwell. Fixed association and Sim(3) composition link feed-forward windows without per-video fitting; Appendix B gives the matched runtime breakdown.




Figure 7: **Crowded in-the-wild reconstruction** on 3DPW `downtown_runForBus`. Top: input and persistent tracks. Bottom: metric scene, camera trajectory and bodies colored by identity. The sequence contains 14 tracks over 10 s, with 10 to 11 people reconstructed per frame.

Table 5: **Association on 13 two-person 3DPW test sequences.** Thresholds tuned on four validation sequences.

| method | IDF1↑ | HOTA↑ | ID switches↓ | fragments↓ | MOTA↑ |
|---|---|---|---|---|---|
| Human3R (native) | **87.4** | 72.3 | 1682 | 87 | 87.1 |
| motion + confidence DP | 66.9 | 56.9 | 97 | 77 | 87.5 |
| raw VGGT query/key | 51.8 | 53.4 | 8065 | 861 | 51.7 |
| projected identity | 60.2 | 54.5 | 664 | 284 | 73.7 |
| **projection + motion + confidence** | 82.1 | **74.5** | **39** | **73** | **96.5** |

**Temporal association.** Table 5 compares association cues on 13 two-person 3DPW test sequences with fixed Multi-HMR proposals and bodies. The combined cost reaches 74.5 HOTA and 39 ID switches, versus 72.3 and 1682 for native Human3R. Human3R leads IDF1 (87.4 versus 82.1) and eight of the 13 sequences; one crowded clip dominates its switch count. The weaker single-cue variants support combining projected identity with motion, but do not isolate confidence. This controlled component test does not measure the final end-to-end pipeline (Appendix A).

**Scale-training stages.** Table 3 tests web-video pseudo-labeling within the final recipe. Removing it while keeping BEDLAM2 (Tesch et al., 2025) initialization and exact RICH/3DPW fine-tuning worsens WA/W/RTE from 66.3/193.6/0.9 to 116.3/255.9/3.7, so the pseudo-label stage contributes beyond exact labels. On held-out RICH, the pseudo-labels have a 14.1% median scale error; the readout pretrained on them reaches 8.4%, and 7.2% after exact fine-tuning (Appendix D).

## 5 Conclusion

We asked whether one up-to-scale 3D foundation model can supply what joint people-scene reconstruction lacks: metric scale and persistent person identity. For scale, people in unlabeled web video proved usable supervision: body-ruler pseudo-labels pretrain a Scale Readout that, after exact metric fine-tuning, predicts scale from the model's own tokens with no ruler or visible person at inference. For identity, the pretrained backbone already links moving people across frames in its intermediate query-key features, although moving people are excluded from its matching supervision, and a small projection turns this into an association cue. With these readouts, *WildHSR* reconstructs people, cameras and scene in one metric world and is the first feed-forward method in our comparison to beat the best optimization-based WA-MPJPE and RTE on EMDB-2. More broadly, 3D foundation models encode more than their explicit outputs, and people in ordinary video can teach them what those outputs lack.

**Limitations.** Identity weakens under interaction and occlusion, and its specificity to people is untested. RICH training mixtures may differ across methods.

## REPRODUCIBILITY STATEMENT

**Models and adaptation.** VGGT-Ω and Multi-HMR retain their released base weights; small LoRAs adapt them (Hu et al., 2022). The backbone LoRA has rank 16, $\alpha = 32$ and dropout 0.05 on the last four frame-attention and last four global-attention blocks (2.10M parameters); the mesh adapter has 0.75M parameters. The Scale Readout uses a 512-dimensional query, four decoder layers and eight heads, trained with AdamW at $10^{-4}$ and batch size 16. Appendix B lists every module, parameter count and training setting.

**Training data and splits.** The Scale Readout is initialized on BEDLAM2, pretrained on offline ruler pseudo-labels from 100,000 curated web clips (about 300 hours), then fine-tuned with exact metric supervision from the standard RICH and 3DPW training splits. Standard validation partitions are used for model selection; EMDB-2 and the RICH and 3DPW test partitions are held out. Section 3.3 gives the ruler equation and the agreement gate, and Appendix D measures the pseudo-labels against held-out metric ground truth. The ruler, agreement gate, ViTPose++-H and 4DHumans are used only to make training labels and are absent at inference.

**Inference and evaluation.** Inference uses 100-frame windows with stride 50; Section 3.4 and Appendix I give the complete composition and association procedure, including every threshold. EMDB-2 and RICH follow the published protocols and evaluators of the compared methods (Appendix C). Appendix A describes the controlled association experiment of Table 5, and Appendix B the runtime measurement. Code, trained models and per-sequence outputs will be released.

## ETHICS AND WEB-DATA STATEMENT

**Data.** Evaluation and exact metric supervision use established public research datasets (EMDB, RICH, 3DPW and BEDLAM2) under their licences. Publicly accessible web video supplies 100,000 clips that are used only to compute non-identifying scale pseudo-labels; we do not redistribute the video. We did not record new video or recruit participants.

**Risks and intended use.** Persistent metric reconstruction of people could support surveillance or tracking of individuals. Track IDs in our system are per-video indices and do not identify people; the system performs no face or identity recognition, and it should not be used for biometric identification or high-stakes decisions about individuals. Web video and parametric body priors may introduce demographic, body-shape, clothing, mobility and visibility biases, and accuracy for under-represented groups has not been measured separately.

## AI USE STATEMENT

Large language models assisted with manuscript drafting, restructuring and editing, and with writing code for evaluation scripts and figures. All reported numbers come from executed code, not from model output. The authors directed this work, checked and revised the text and code, and take full responsibility for this paper.

## REFERENCES


Keito Abe, Kaede Shiohara, Takashi Otonari, and Toshihiko Yamasaki. GUSH3R: Everyone everywhere all at once as gaussians. *arXiv preprint arXiv:2607.05243*, 2026.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

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 *ECCV*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3R: Estimating disentangled motion from DUSt3R without training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: Everyone everywhere all at once. In *International Conference on Learning Representations (ICLR)*, 2026.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhongyuan Hu, Qijun Ying, Jiazhi Shu, Ronghui Li, Yu Lu, Zijiao Zeng, and Xiu Li. WATCH: World-aware allied trajectory and pose reconstruction for camera and human, 2025.

Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan Zarate, and Otmar Hilliges. EMDB: The electromagnetic database of global 3d human pose and shape in the wild. In *ICCV*, 2023.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*, 2026.

Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. PACE: Human and camera motion estimation from in-the-wild videos. In *3DV*, 2024.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

Jiefeng Li, Ye Yuan, Davis Rempe, Haotian Zhang, Pavlo Molchanov, Cewu Lu, Jan Kautz, and Umar Iqbal. COIN: Control-inpainting diffusion prior for human and camera motion estimation. In *European Conference on Computer Vision (ECCV)*, 2024.

Mengfei Li, Peng Li, Zheng Zhang, Jiahao Lu, Chengfeng Zhao, Wei Xue, Qifeng Liu, Sida Peng, Wenxiao Zhang, Wenhan Luo, Yuan Liu, and Yike Guo. UniSH: Unifying scene and human reconstruction in a feed-forward pass. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14038–14049, 2026.

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4D human-scene reconstruction in the wild. In *International Conference on Learning Representations (ICLR)*, 2026.

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):248:1–248:16, 2015.

Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019.

Sara Rojas, Matthieu Armando, Bernard Ghanem, Philippe Weinzaepfel, Vincent Leroy, and Gregory Rogez. HAMSt3R: Human-aware multi-view stereo 3D reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia*, 2024.

Boao Shi, Qiao Feng, Yiming Huang, and Lingjie Liu. Scene and human in one world: Reconstruction in a feedforward pass, 2026.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*, 2024.

Chentao Song, He Zhang, Haolei Yuan, Haozhe Lin, Jianhua Tao, Hongwen Zhang, and Tao Yu. MetricHMSR: Metric human mesh and scene recovery from monocular images. *arXiv preprint arXiv:2506.09919*, 2025.

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 *CVPR*, 2023.

Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J. Black. BEDLAM2.0: Synthetic humans and cameras in motion. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

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 *European Conference on Computer Vision (ECCV)*, pp. 601–617, 2018.

Hengyi Wang and Lourdes Agapito. AMB3R: Accurate feed-forward metric-scale 3d reconstruction with backend. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Yifan Wang, Brian L. Curless, and Steven M. Seitz. People as scene probes. In *European Conference on Computer Vision (ECCV)*, 2020.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3d humans from in-the-wild videos. In *ECCV*, 2024b.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2): 1212–1230, 2024.

Fengyuan Yang, Kerui Gu, Ha Linh Nguyen, Tze Ho Elden Tse, and Angela Yao. Humans as checkerboards: Calibrating camera motion scale for world-coordinate human mesh recovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Jiahui Yang, Donglin Di, Xuancheng Zhang, Lei Fan, Jianxun Cui, Hao Li, and Baorui Ma. MetricAnything: Scaling metric depth pretraining with noisy heterogeneous sources. In *European Conference on Computer Vision (ECCV)*, 2026.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023.

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations (ICLR)*, 2025.