TRAM: Global Trajectory and Motion of 3D Humans from in-the-wild Videos

Yufu Wang[1], Ziyun Wang[1], Lingjie Liu[1], and Kostas Daniilidis[1,2]

[1] University of Pennsylvania
[2] Archimedes, Athena RC

Abstract. We propose TRAM, a two-stage method to reconstruct a human's global trajectory and motion from in-the-wild videos. TRAM robustifies SLAM to recover the camera motion in the presence of dynamic humans and uses the scene background to derive the motion scale. Using the recovered camera as a metric-scale reference frame, we introduce a video transformer model (VIMO) to regress the kinematic body motion of a human. By composing the two motions, we achieve accurate recovery of 3D humans in the world space, reducing global motion errors by a large margin from prior work. https://yufu-wang.github.io/tram4d/

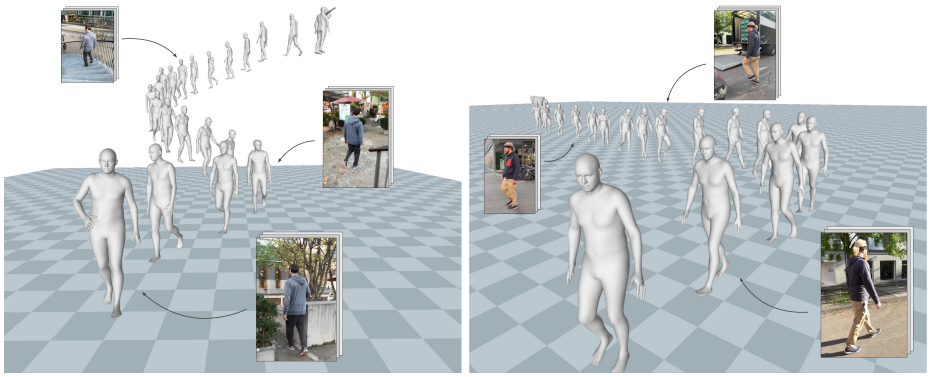

Fig. 1: Overview. Given an in-the-wild video, TRAM reconstructs the complete 3D human motion: **global trajectory and local body motion**, in diverse and long-range scenarios.

1 Introduction

Understanding human movement from visual data is a fundamental problem in computer vision. To truly see how we move, we need to capture not only the kinematic body poses but also our global trajectory in the world space as in Fig.1. This task is challenging when the input is a casual video where the human is observed from a moving camera. Without reasoning about the camera motion, most prior studies reconstruct humans in the camera frame, making it difficult to contextualize a human's action in the environment. The ability to capture this complete motion from videos will unlock many applications, such

as reasoning about our interaction with the scene [56, 101], imitation learning of life-like behaviors [61, 64, 74], gait analysis with metric-scale measurement for health care [32, 75], and beyond.

Our observation is that if we can precisely locate the camera trajectory **in the world frame**, and estimate a person's body motion **in the camera frame**, then we can recover the person's complete motion in the world by composing the two. This two-stage approach naturally disentangles the motion estimation problem into two well-defined subproblems. The goal of this paper is to address the two subproblems so that when combined they can capture accurate human motion in the world from a video.

For camera trajectory estimation, Simultaneous Localization and Mapping (SLAM) [9] has been widely used in Robotics. However, traditional SLAM faces two key challenges in our context. Firstly, SLAM assumes a static environment while our videos contain moving humans, which decreases the estimation accuracy. Secondly, monocular SLAM only recovers camera trajectory up to an arbitrary scale. To represent camera movement in the metric world frame, we need to estimate a scaling factor. Recent studies propose to infer the scale of camera motion from the observed human motion [22, 95]. Specifically, after SLAM, these methods jointly optimize human poses and the camera scale, so that the human displacements match a learned locomotion model. However, because the motion models are learned from studio MoCap data, the predicted displacement does not generalize to real-world complexity. Consequently, they struggle to recover complex and long-range trajectories.

In this work, we propose to robustify and metrify SLAM's camera estimation without relying on a human motion model. To make it robust to dynamic humans, we use a masking procedure to dually remove the dynamic regions in the input images and the dense bundle adjustment steps. Enforcing SLAM to only use the background for camera estimation from the beginning reduces the chance of catastrophic failure. To convert the camera estimation to metric scale, we utilize semantic cues from the background. Our insight is that the environment provides rich and reliable information about the scale of the scene. When we watch Spider-Man swing across skyscrapers, we understand the distance he travels despite the lack of locomotion: the environment serves as a scale. We demonstrate that this information can be derived, in a reliable manner, from noisy depth predictions [5]. As a result, we recover accurate and metric-scale camera motion to serve as a reference frame for the local human motion.

The second challenge is the recovery of the kinematic body motion in the camera frame. While seeing great progress from deep regression networks [28, 37], the de facto per-frame prediction lacks temporal coherence. However, video-based models fall behind in accuracy compared to per-frame models. The main bottleneck in developing a scalable video model is the training cost with videos (concerning larger models) and the lack of video data (compared to the images). To address this issue, our approach leverages a large pre-trained model that has learned rich representations of the human body, and finetunes it on video data.

We propose Video Transformer for Human Motion (VIMO), that builds on top of the large pre-trained HMR2.0 [17]. VIMO adds two temporal transformers, one to propagate temporal information in the image domain, and the other to propagate information in the motion domain. We keep the ViT-Huge backbone frozen to preserve the learned representations, and finetune the two transformers and the original transformer decoder on videos. Without bells and whistles, this fully "transformerized" design achieves state-of-the-art reconstruction accuracy.

Our method takes a scene-centric point of view in estimating human movements, versus recent methods with a human-centric point of view using motion priors. By solely using the scene background to estimate metric scale camera motion, and always reconstructing humans in the camera frame, our two-stage approach recovers the global trajectory and motion (TRAM) of 3D humans from videos, with a large error reduction in global trajectory from the best published results.

Contributions. (i) We propose a general method TRAM, that recovers human trajectory and motion from in-the-wild videos with a large improvement over prior works. (ii) We demonstrate that the human trajectory can be inferred from SLAM, and provide technical solutions to make monocular SLAM robust and metric-scale in the presence of dynamic humans. (iii) We propose the video transformer model VIMO that builds on top of a large pre-trained image-based model, and demonstrate that this scalable design achieves state-of-the-art pose and shape reconstruction.

2 Related Work

3D Human Recovery from Images and Videos. Reconstructing 3D human is most widely formulated as recovering the parameters of a parametric human model, such as SMPL [48] and its followups [58, 65, 66, 92]. Without reasoning about the camera movement, the majority of the works consider reconstruction in a local coordinate system, often the camera coordinate frame or a multi-view reference frame [26, 28, 37, 47, 77, 79, 100].

In the optimization paradigm, the SMPL model can be fitted through energy minimization to images [1, 7, 15, 71], videos [3, 24, 91], and other sensing modalities [83, 90]. In contrast, regression approaches learn pose and shape representation from large datasets [25, 45, 52], and predict the parameters given a single image or video frames [28, 34, 55]. The continuous improvement of accuracy stems, partly from the improvement of datasets [6, 27, 83], and partly from ingenious designs that imbue domain knowledge into neural architectures [10, 12, 38, 43, 44, 53]. Notable designs includes reprojection alignment [41, 99], part-based reasoning [35, 40, 84], temporal pooling [11, 29, 50, 89], and learning-based optimization [13, 73, 88, 97].

In contrast, HMR2.0 [17] demonstrates that a general-purpose transformer architecture [16, 82], with enough capacity and data, can learn robust representations that in many cases outperform domain-specific designs. In this work, we turn HMR2.0 into a video model with the same principle, by adding two

transformer encoders to propagate temporal information from nearby frames. We demonstrate that this simple and fully "transformerized" design outperforms existing video models for 3D human regression.

3D Human Recovery in the World Space. Reconstructing a person's complete movement through space (global trajectory and motion) is crucial for understanding our actions in the world. Integrating additional sensors or cameras allows methods to capture global human motion reliably [18, 21, 30, 57, 67, 83].

This recovery is difficult with in-the-wild videos, where the estimation is with respect to a moving camera. Methods such as GLAMR [96] and D&D [39] use a human's locomotion to estimate the global trajectory but are not reliable under real-world complexity. In contrast, [46, 60] densely reconstruct the environment with SfM [68], and contextualize the human placement with human-scene constraints. This approach requires a priori reconstruction of the environment, which is not always possible with in-the-wild videos. TRACE [78] and WHAM [70] propose to regress the per-frame pose and translation directly. They achieve great results by learning to regress trajectories from MoCap data, but its reliance on this data as a prior hinders its ability to predict novel trajectories.

A line of works employs a hybrid approach [22,36,95]. They utilize SLAM [80] for camera trajectory estimation (up to a scale) to initialize the human placement and follow with optimization to determine the human poses and the movement scale. They leverage motion models [19, 63] to derive motion scale, but struggle with complex real-world scenarios like navigating stairs or parkour, which are not adequately captured by MoCap data [51]. Wang et al. [86,87] use a checkerboard to calibrate the motion scale of egocentric videos, but this technique requires objects of exact dimensions to be present.

Our method, TRAM, also recovers camera motion with SLAM, but we differ by deriving the movement scale from the background. By avoiding human motion priors for trajectory estimation, it achieves better generalization to complex scenarios. Notably, we also address SLAM's robustness with dynamic humans.

3 Method

Our goal is to recover the complete 3D human motion from videos in the wild. We decompose this motion into its $SE(3)$ root trajectory $\{\mathbf{H}_t\}_{t=0}^T$ in the world frame and the kinematic body motion $\{\mathbf{\Theta}_t\}_{t=0}^T$ represented by a sequence of SMPL poses in the camera frame.

To infer the human trajectory in the world, our approach is to estimate the camera trajectory $\{\mathbf{G}_t\}_{t=0}^T$ and the human's positions with respect to the camera $\{\mathbf{T}_t\}_{t=0}^T$ at each time step. We demonstrate that by making the camera trajectory estimation robust (sec. 3.2) and metric-scale (sec. 3.3), the human trajectory can be accurately recovered as $\{\mathbf{H}_t\}_{t=0}^T = \{\mathbf{G}_t \circ \mathbf{T}_t\}_{t=0}^T$.

To reconstruct the kinematic body motions, we propose VIMO (sec. 3.4), a video-inference transformer model that reconstructs body motion $\{\mathbf{\Theta}_t\}_{t=0}^T$ and relative positions $\{\mathbf{T}_t\}_{t=0}^T$ in the camera frame.

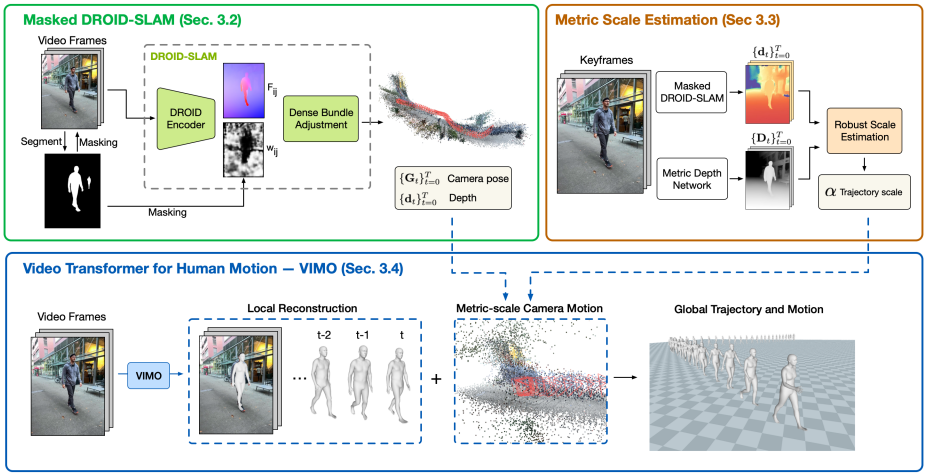

Fig. 2: Overview of TRAM. Top-left: given a video, we first recover the relative camera motion and scene depth with DROID-SLAM, which we robustify with dual masking (Sec. 3.2). Top-right: we align the recovered depth to metric depth prediction with an optimization procedure to estimate metric scaling (Sec. 3.3). Bottom: We introduce VIMO to reconstruct the 3D human in the camera coordinate (Sec. 3.4), and use the metric-scale camera to convert the human trajectory and body motion to the global coordinate.

3.1 Preliminary: 3D Human Model

We use SMPL [48] to represent the 3D human body. The SMPL model is a parametric mesh model $\mathcal{M}(\theta, \beta, r, \pi) \in \mathbb{R}^{6890\times 3}$, where $\theta \in \mathbb{R}^{23\times 3}$ are the relative rotations of the 23 body joints, $\beta \in \mathbb{R}^{10}$ is the shape parameter, and $r \in \mathbb{R}^3$ and $\pi \in \mathbb{R}^3$ are the root orientation and translation w.r.t the camera. In our context of motion in the world frame, $\mathbf{T}_t = \{r_t, \pi_t\}$ represents the relative position, and $\mathbf{\Theta}_t = \{\theta_t, \beta_t\}$ represents the local body pose and shape. The scale of the SMPL mesh is in metric units, representing the real size of the human in meters.

3.2 Masked DROID-SLAM

We utilize DROID-SLAM [80] to recover the camera trajectory $\{\mathbf{G}_t\}_{t=0}^T$ from monocular videos. In this paper, we propose to reduce the negative effect of dynamic objects through a two-step masking.

For a input video frame $\mathbf{I}_i$, DROID first computes the 2D flows $\mathbf{F}_{ij} \in \mathbb{R}^{h\times w\times 2}$ and its confidence $\mathbf{w}_{ij} \in \mathbb{R}^{h\times w\times 2}$ with respect to nearby keyframes $\{\mathbf{I}_j\}$. Then the dense bundle adjustment layer (DBA) optimizes the relative camera pose $\mathbf{G}_{ij} \in SE(3)$ and depth of the current frame $\mathbf{d}_i \in \mathbb{R}^{h\times w}$ by solving the following objective with Gauss-Newton.

$$E(G, d) = \sum_{(i,j)} \| p_{ij} - \Pi(G_{ij} \circ \Pi^{-1}(p_i, d_i)) \|^2_{\sum_{ij}} \sum_{ij} = \text{diag}\mathbf{w}_{ij} \quad (1)$$

where $\mathbf{p}_i$ are the pixel coordinates and $\mathbf{p}_{ij} = \mathbf{p}_i + \mathbf{F}_{ij}$. This objective minimizes the flow reprojection error weighted by confidence. Global bundle adjustment and loop-closure are applied after all frames are processed.

The predicted confidence $\mathbf{w}_{ij}$ allows DROID to down-weight correspondences with high uncertainty from the DBA process, making it robust to small moving objects. However, when the dynamic object occupies a larger area, the predicted confidence may not be accurate. Pixel coordinates that violate the static scene assumption degrade the accuracy of camera motion estimation.

We propose to dually mask the input image $\hat{\mathbf{I}}_i = \text{mask}(\mathbf{I}_i)$ and the predicted confidence $\hat{\mathbf{w}}_{ij} = \text{mask}(\mathbf{w}_{ij})$, setting the dynamic region to value zero as shown in Fig.2. Masking the flow confidence $\mathbf{w}_{ij}$ is equivalent to removing the dynamic object coordinates from the calculation of reprojection error in Eq.1. This step ensures the DBA only uses background pixels to estimate camera motion, making it robust to large moving objects. Additionally, we also find it beneficial to mask the input images. DROID's learned encoder uses both local and global features to predict dense flow on the image plane. The global feature provides useful context for flow estimation, but large moving objects may contribute motion cues that negatively impact other regions. Masking the input image helps to mitigate this effect on the global feature. We can predict accurate masks of dynamic objects using an object detector [85] and Segment Anything [33]. In this paper, we focus on humans, but other categories can also be detected and masked out.

Our experiments show that the two masking steps are essential, preventing DROID-SLAM from diverging in many sequences where moving humans occupy a large foreground. Masked DROID allows us to recover accurate camera motion, from which we can then infer the human's trajectory as seen from the camera.

3.3 Trajectory Scale Estimation

Masked DROID can recover camera trajectory and scene structure up to an arbitrary scale, but we need them to be in metric scale to represent motion in the world frame and be compatible with SMPL. We show that a scaling parameter that indicates the actual size of the scene can be solved from the semantic of the scene.

Common objects such as buildings, cars, and trees are of known sizes, and as humans we have a mental model that reasons about their spatial scale and relationship. This spatial reasoning capacity is partially presented in a depth prediction network. We can use metric depth prediction to reason about the scale of the camera motion and the scene reconstructed by SLAM. In this paper, we use ZoeDepth [5] to predict metric depth for video frames. Given a keyframe $\mathbf{I}_i$, ZoeDepth predicts the scene depth $\mathbf{D}_i \in \mathbb{R}^{h \times w}$ in meters. DROID returns the depth $\mathbf{d}_i$ in a random unit. By solving for a scaling term α that aligns $\alpha * \mathbf{d}_i$ to $\mathbf{D}_i$, we can re-scale the reconstructed scene and the camera trajectory to the metric unit.

However, depth prediction is usually noisy. The network can under or over-estimate the depth in certain regions of the image. For example, we find ZoeDepth

to often underestimate the distance in the sky region. Similarly, the depth predictions are not temporally consistent and can be inaccurate in many frames. Therefore, we need to derive the scale robustly. To mitigate the effect of bad depth prediction at certain frames, we will solve for the scale for each keyframe independently and take the median over all the keyframes. To solve for the scale for each keyframe, we minimize the energy with robust least square as

$$E(\alpha) = \sum_{(h,w)} \rho(\alpha * \mathbf{d}_i - \mathbf{D}_i) \tag{2}$$

where ρ is the German-McClure robust loss function with the summation over the entire image. We use BFGS for the optimization. Furthermore, we find that depth predictions are less accurate in the far region such as the sky or building in the far distance. Therefore, we set thresholds to exclude the far region from the optimization, essentially using only the middle region where depth prediction is more reliable to solve for the scale.

The thresholding and robust least squares handle noisy areas in a frame, and the median allows us to use the statistics of the whole trajectory. Overall, our experiments show that this procedure provides a good estimate for the scale of the scene, and is more reliable than inferring scale from human motion models.

3.4 Video Transformer for Human Motion

Human pose and shape regression networks can now reconstruct 3D people from a single image in the general case, but videos provide temporal constraints to help reconstruct motions that look natural and smooth. In this work, we propose **VI**deo Transformer for Human **MO**tion (VIMO) to accomplish this goal.

A recent success recipe in vision is to finetune a large pre-trained model for downstream tasks. HMR2.0 [17] demonstrates state-of-the-art reconstruction by building on top of a pre-trained ViT-H [20, 93] and scaling up its training data and compute. Our design philosophy for VIMO is therefore twofold: we aim to utilize the rich knowledge in a large pre-trained model, and make flexible and scalable architecture design whose performance can scale with data.

Toward this goal, we turn the image-based HMR2.0 into a video model by adding two temporal transformers, as shown in Fig. 3. HMR2.0 uses a ViT-H to extract image features and a standard transformer decoder with a zero token to cross-attend the image features and regress the final outputs. The two new temporal transformers from VIMO use the encoder-only architecture but distinctively propagate temporal information in the image domain and the human motion domain.

The first transformer applies attention across time on each patch token from ViT. This operation is repeated at different spatial locations independently. The combination of ViT and this temporal transformer can be considered as a factorized spatial-temporal model [2], where the ViT applies attention across space and the temporal transformer applies attention across time. The role of this layer is to use temporal correlation, such as appearance and motion cues, to make the image features more accurate and robust.

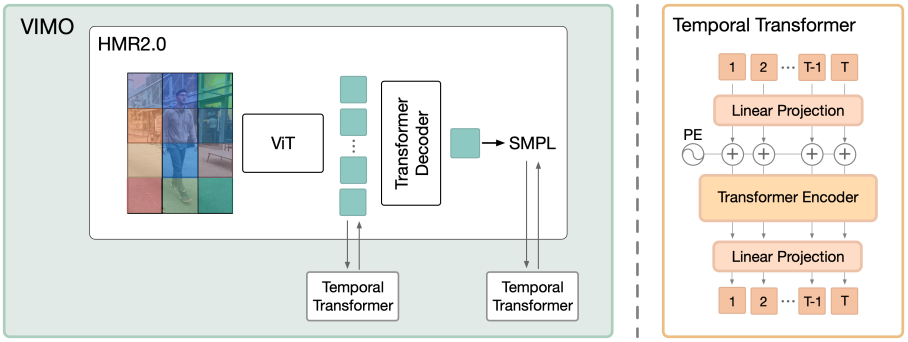

Fig. 3: Video transformer VIMO builds on top of the large pre-trained HMR2.0 and adds two temporal transformers to propagate information across video frames. Right: the temporal transformers use the same encoder-only architecture. ▢ represents patch tokens at the same spatial location across time in the first temporal module, and represents SMPL poses across time in the second temporal module. More details are included in the supplementary.

The second transformer encodes and decodes a sequence of SMPL poses. The goal here is to learn a prior on the human motion space so that noisy poses can be corrected to become smooth motion. Previous work largely learns temporal models in the global feature space [34, 50, 59]. We conjecture that the latent space before the regressor is sub-optimal to learn a motion model, as this space entangles many other information such as shape, camera, and image features. In contrast, studies in motion generation and denoising [4, 62] show that motion, represented as a sequence of poses, can be directly encoded and decoded with transformers. Therefore, we apply this general architecture on SMPL poses $\{\theta_t, r_t\}$ directly. While such a motion model can be pre-trained with MoCap data, we demonstrate end-to-end learning from videos.

We keep the ViT backbone frozen and finetune VIMO on video data with the following losses.

$$\mathcal{L} = \lambda_{2D}\mathcal{L}_{2D} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_{SMPL}\mathcal{L}_{SMPL} + \lambda_{V}\mathcal{L}_{V} \tag{3}$$

Each term is calculated as

$$\begin{aligned}
\mathcal{L}_{2D} &= ||\hat{\mathcal{J}}_{2D} - \Pi(\mathcal{J}_{3D})||_F^2 \\
\mathcal{L}_{3D} &= ||\hat{\mathcal{J}}_{3D} - \mathcal{J}_{3D}||_F^2 \\
\mathcal{L}_{SMPL} &= ||\hat{\Theta} - \Theta||_2^2 \\
\mathcal{L}_{V} &= ||\hat{V} - V||_F^2
\end{aligned}$$

where $\mathcal{J}_{3D}$ and V are the 3D joints and vertices obtained from the SMPL model, and the hat operator denotes the ground truth of that variable. Our experiments show that both temporal transformers are essential for VIMO to reconstruct accurate and smooth motion from videos.

4 Experiments

Datasets. We use three datasets to train our video model VIMO: 3DPW [83], Human3.6M [25], and BEDLAM [6]. As models are trained on increasingly diverse data, we create a baseline by finetuning HMR2.0 on the same three datasets as a fair comparison to validate our design. We evaluate human motion and shape reconstruction on 3DPW and EMDB (subset 1) [30]. For human trajectory recovery, we evaluate on the EMDB dataset (subset 2), which contains 25 sequences with ground truth for both human and camera trajectory.
Implementation. We train VIMO for 100K iterations using AdamW [49] with weight decay of 0.01 and a batch size of 24 sequences, each with a 16-frame window. We use a learning rate of 1e-5 for the transformer decoder and 3e-5 for the two temporal transformers. We include more details about the architecture in the supplementary. For masking, we use detections from YOLOv7 [85] as the prompt for the Segment Anything Model [33]. We use ZoeDepth [5] to predict metric depth from videos.
Evaluation metrics. For pose and shape reconstruction, we use the common reconstruction metrics: mean per-joint error (MPJPE), Procrustes-aligned per-joint error (PA-MPJPE), and per-vertex error (PVE). To evaluate the motion smoothness, we compute acceleration error (ACCEL) against the ground truth acceleration.

For human trajectory evaluation, we slice a sequence into 100-frame segments and evaluate 3D joint error after aligning the first two frames (W-MPJPE$_{100}$) or the entire segment (WA-MPJPE$_{100}$) [95]. Following WHAM [70], we evaluate root translation error normalized by the total displacement (RTE in %) after rigid alignment without scaling. In addition, we compute egocentric-frame root velocity error (ERVE) to measure the root motion accuracy.

For camera trajectory evaluation, we follow SLAM literature and evaluate absolute trajectory error (ATE) [76], which performs rigid alignment with scaling to align the camera trajectory with ground truth before computing error. To evaluate the accuracy of our scale estimation, we also evaluate ATE using our estimated scale (ATE-S) [36].

4.1 Comparison on Camera Trajectory Recovery

Dynamic masking. We first validate the two-step masking procedure by evaluating the camera trajectory accuracy on EMDB. We create two baselines with ORB-SLAM2 [54] and DROID-SLAM [80].

As shown in Table 1, the default DROID-SLAM has a large error on EMDB due to dynamic humans. Compared to DROID, ORB-SLAM2 is less affected by dynamic humans, likely due to its outlier rejection procedure (e.g. RANSAC) and only using sparse points. However, ORB-SLAM2 loses track in 9 out of 25 sequences due to the sudden loss of feature points. Masking improves the results of both methods, particularly in longer sequences.

Masking humans in the input images improves the accuracy. The improvement is more significant with masking in the dense bundle adjustment process.

Methods	EMDB 2 (ATE)			
	Short(5)	Medium(10)	Long(10)	Average
ORB-SLAM2	0.08	0.29	2.08	1.05
ORB-SLAM2 + Mask Image	0.38	0.60	1.10	0.79
DROID-SLAM	0.40	2.55	3.31	2.42
DROID-SLAM + Mask Image	0.36	0.63	2.74	1.42
DROID-SLAM + Mask DBA	0.45	0.42	1.63	0.91
Masked DROID	**0.32**	**0.20**	**0.44**	**0.32**

Table 1: Evaluation of camera estimation **with ground truth scale (ATE)**. Results are grouped according to sequence length: short(<20m), medium(<60m) and long(>60m). Parenthesis denote the number of sequences. ORB-SLAM2 fails in 9/25 sequences so its results are calculated with the other 16 sequences. ATE is in m.

Methods	EMDB 2 (ATE-S)			
	Short(5)	Medium(10)	Long(10)	Average
Masked DROID + ZoeDepth	**0.33**	2.29	5.27	3.09
Masked DROID + Scale est.	0.48	**0.62**	**0.78**	**0.66**

Table 2: Evaluation of camera estimation **with estimated scale (ATE-S)**. Naively using ZoeDepth predictions as depth input for DROID results in large error. The proposed method produces good scale estimation. ATE-S is in m.

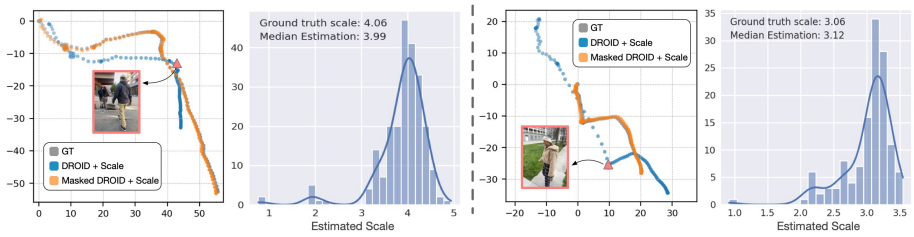

Fig. 4: Camera trajectory estimation. With dynamic humans in the scene, the default DROID-SLAM tends to diverge. The two-step masking makes it robust. In addition, our procedure estimates a reasonable metric scale for the cameras.

By masking both the input and the DBA, we achieve the largest improvement over the baseline. We observe that all previously diverged sequences now have a more reasonable trajectory, as shown in Figure 4.

Metric scaling. The above evaluation calculates ATE with the ground truth scale, but our target application requires us to estimate the scale. We evaluate this error in Table 2. As a baseline, we give ZoeDepth prediction directly to DROID using its RGB-D mode. This naive combination causes large errors because incorrect depth predictions lead the bundle adjustment to wrong solutions. Currently, metric depth prediction is not accurate enough to replace RGB-D inputs. Our method circumvents noisy depth prediction with robust optimization and uses the median of the whole trajectory. Compared to using the ground

Models	EMDB 2				
	PA–MPJPE	WA-MPJPE$_{100}$	W-MPJPE$_{100}$	RTE	ERVE
TRACE [78]	58.0	529.0	1702.3	17.7	370.7
GLAMR [96]	56.0	280.8	726.6	11.4	18.0
SLAHMR [95]	61.5	326.9	776.1	10.2	19.7
WHAM (w/ DROID) [70]	38.2	133.3	343.9	4.6	14.7
TRAM	**38.1**	**76.4**	**222.4**	**1.4**	**10.3**

Table 3: Evaluation of human global trajectory and motions. RTE is in %, ERVE is in mm/frame, and the other pose metrics are in mm. Ordered by RTE.

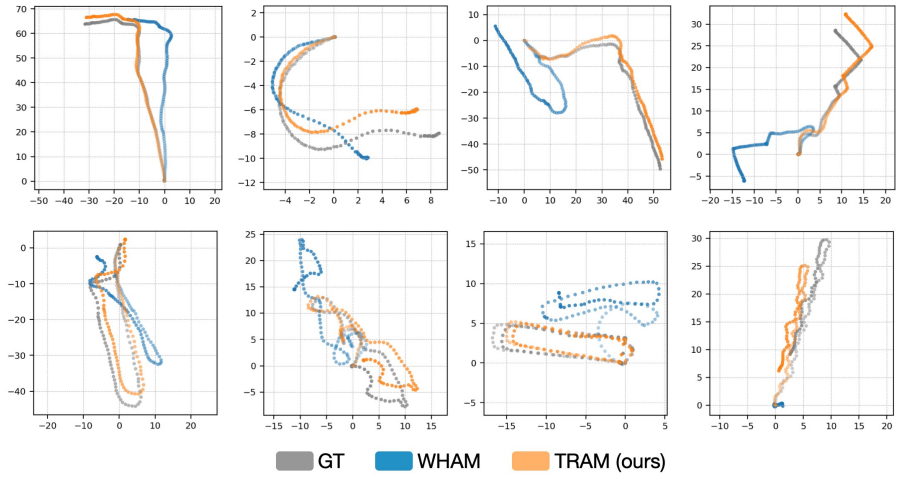

Fig. 5: Human global trajectory on EMDB. Compared to WHAM, our method produces less drift and a more accurate scale for complex and long-range tracks.

truth scale, it increases the error only by 30cm, a gap that will likely become smaller with continuous improvement of depth prediction [31, 94].

We visualize the distribution of the estimated scales among keyframes in a sequence in Figure 4. As shown, not all frames are suitable for estimating scale due to noisy or incorrect depth prediction. Depth prediction can also incur a large bias in frames where there is a lack of references. Using the median of the whole sequence approximates the ground truth scale well. Figure 4 shows two sequences in which DROID diverges without masking. Our method estimates accurate and metric-scale camera trajectory, which becomes a reference frame when reasoning about the human motion in the world.

4.2 Comparison on Human Trajectory Recovery

With accurate and metric-scale camera trajectory, we can infer the human trajectory as proposed in Section 3, using VIMO to reconstruct the 3D human with

		3DPW (14)				EMDB (24)			
	Models	PA-MPJPE	MPJPE	PVE	Accel	PA-MPJPE	MPJPE	PVE	Accel
per-frame	SPIN [37]	59.2	96.9	112.8	31.4	87.1	140.7	166.1	41.3
	PARE [35]	46.5	74.5	88.6	–	72.2	113.9	133.2	–
	CLIFF [41]	43.0	69.0	81.2	22.5	68.3	103.3	123.7	24.5
	HybrIK [40]	41.8	71.6	82.3	–	65.6	103.0	122.2	–
	HMR2.0 [17]	44.4	69.8	82.2	18.1	60.7	98.3	120.8	19.9
	ReFit [88]	40.5	65.3	75.1	18.5	58.6	88.0	104.5	20.7
temporal	TCMR [11]	52.7	86.5	101.4	6.0	79.8	127.7	150.2	5.3
	VIBE [34]	51.9	82.9	98.4	18.5	81.6	126.1	149.9	26.5
	MPS-Net [89]	52.1	84.3	99.0	6.5	81.4	123.3	143.9	6.2
	GLoT [69]	50.6	80.7	96.4	6.0	79.1	119.9	140.8	5.4
	GLAMR [96]	51.1	–	–	8.0	73.8	113.8	134.9	33.0
	TRACE [78]	50.9	79.1	95.4	28.6	71.5	110.0	129.6	25.5
	WHAM (ViT) [70]	35.9	**57.8**	**68.7**	6.6	50.4	79.7	94.4	5.3
	HMR2.0(ft)	37.3	63.2	74.3	14.8	49.7	82.7	95.3	20.5
	VIMO	**35.6**	59.3	69.6	**4.9**	**45.7**	**74.4**	**86.6**	**4.9**

Table 4: Comparison of mesh reconstruction on the 3DPW and EMDB datasets. HMR2.0(ft) is our baseline by finetuning HMR2.0b on the same training data as VIMO. Parenthesis denotes the number of body joints used to compute errors for the dataset. Bold numbers denote the best performance. Accel is in m/s^2, and others are in mm.

respect to the camera frame. As shown in Table 3, we improve global pose and trajectory accuracy by a large margin. Particularly, we achieve a 60% error reduction in the root trajectory estimation (RTE). This confirms our observation that accurate camera recovery is essential in estimating human motion in the world frame.

We visualize the human trajectory in the xy-plane in Figure 5. As observed, WHAM can recover good trajectories when the humans walk in mostly straight lines. If the trajectories involve large curves, WHAM's learned regressor cannot recover an accurate trajectory. In contrast, we recover complex trajectories with a minimum drift and a more accurate scale. Moreover, WHAM fails when the motion is outside the MoCap data, such as walking down stairs or riding a skateboard, as shown in Figure 6. We can recover such complex trajectories because our method does not depend on a learned prior from MoCap data.

4.3 Comparison on Human Body Motion Reconstruction

We evaluate human mesh reconstruction from VIMO in Table 4. Without domain-specific designs, VIMO outperforms all other methods in both reconstruction accuracy and motion smoothness.

To get a fair comparison, we create a baseline HMR2.0(ft), by finetuning the transformer decoder of HMR2.0 with the same training data and procedure as

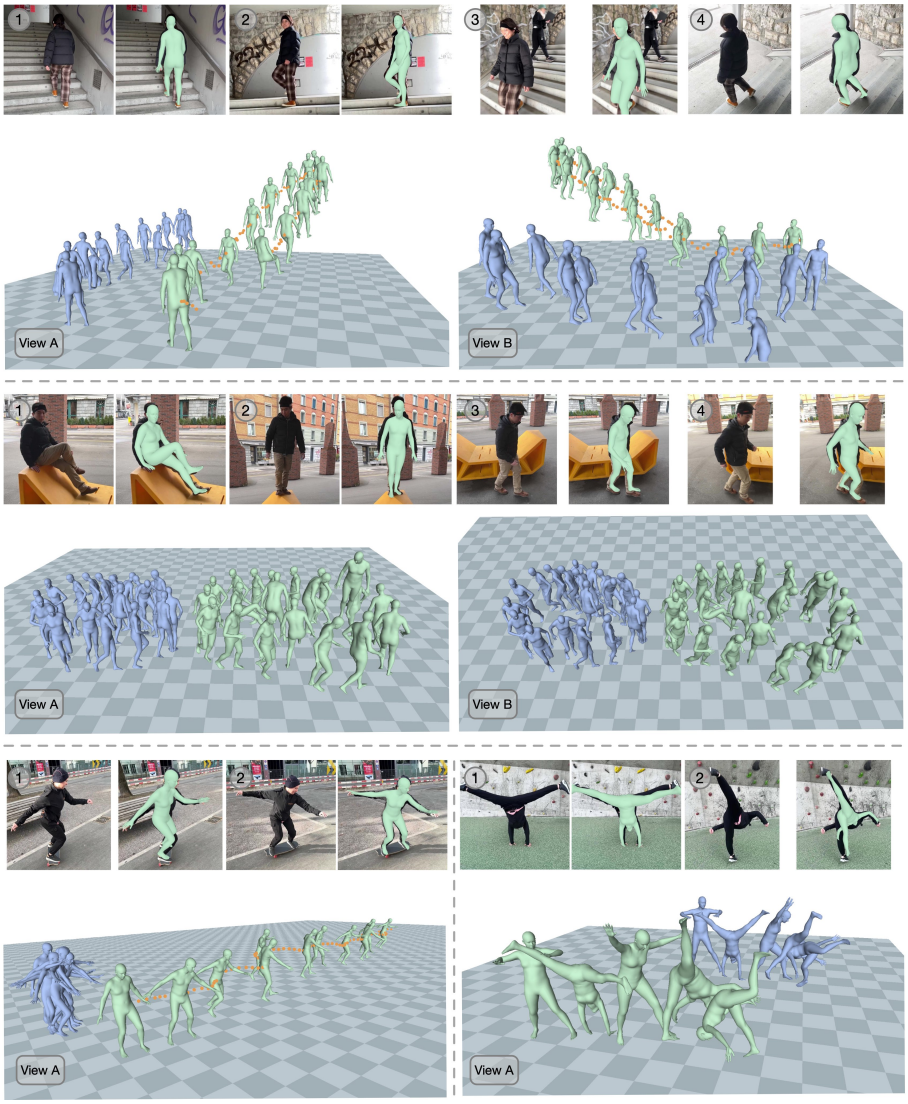

Fig. 6: Human motion and trajectory: WHAM ■ vs Ours ■. We produce more accurate motion and tracks that generalize to diverse terrain and motion complexity. WHAM's results are from the version that uses ground truth gyro.

VIMO. HMR2.0(ft) has a higher benchmark performance than the official release because it uses additional data from 3DPW and BEDLAM. VIMO achieves consistent improvements over the baseline, showing the effectiveness of the two temporal transformers.

We conduct ablations on VIMO. As shown in Table 5, both temporal transformers are important. Removing the token temporal transformer decreases the

	3DPW (14)				EMDB (24)			
Models	PA-MPJPE	MPJPE	PVE	Accel	PA-MPJPE	MPJPE	PVE	Accel
HMR2.0(ft)	37.3	63.2	74.3	14.8	49.7	82.7	95.3	20.5
+ Tokens Attention	36.3	59.5	**69.5**	8.9	**45.4**	**74.1**	**85.6**	11.6
+ Motion Attention	37.0	60.5	71.1	4.9	48.4	78.1	89.2	5.2
VIMO	**35.6**	**59.3**	69.6	**4.9**	45.7	74.4	86.6	**4.9**

Table 5: Ablation on VIMO. Removing either temporal transformer decrease reconstruction accuracy or motion smoothness. The proposed VIMO recovers accurate and smooth motion.

reconstruction accuracy, indicating that there is information gain in the image feature domain by considering neighboring frames with attention. Removing the motion transformer decreases motion smoothness, as it plays a key role in denoising to produce smooth and natural motion.

We provide more qualitative results in Figure 6. VIMO's reconstruction aligns well with the input frames and handles complex poses gracefully. We attribute this robustness to its pre-training from HMR2.0. Freezing the ViT-Huge backbone preserves the recognition power. Since the whole VIMO architecture is made up of transformers, this is a general and scalable design that can take advantage of scaling the model size, compute, and data [14, 42, 98].

4.4 Limitations

While TRAM works well on datasets, we detail the following challenges for future research. Firstly, SLAM's dependence on known focal length limits its applicability in many in-the-wild cases. Future research will need to address the estimation of focal length during bundle adjustment [8, 72]. Secondly, depth estimation tends to be less accurate with extreme focal length, and methods that consider focal length will be beneficial [23]. Lastly, our method follows a strict separation of camera and human motion but a joint optimization in the end could be beneficial [36]. Joint optimization with proper physics priors will improve implausible movements such as foot sliding and penetration [81].

5 Conclusions

We presented TRAM, a new two-stage method to recover the global human trajectory and body motion from in-the-wild videos with moving cameras. TRAM is efficient and accurate, improving global human reconstruction by large margins. We also introduced VIMO, a video transformer model for regressing the local human body motion. VIMO is simple and scalable, and outperforms prior models in different pose and shape reconstruction benchmarks.

Acknowledgements The authors appreciate the support of the following grants: NSF NCS-FO 2124355, NSF FRR 2220868, NSF IIS-RI 2212433.

References

1. Easymocap - make human motion capture easier. Github (2021), https://github.com/zju3dv/EasyMocap
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
3. Arnab, A., Doersch, C., Zisserman, A.: Exploiting temporal context for 3d human pose estimation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3395–3404 (2019)
4. Baradel, F., Brégier, R., Groueix, T., Weinzaepfel, P., Kalantidis, Y., Rogez, G.: Posebert: A generic transformer module for temporal 3d human modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
5. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
6. Black, M.J., Patel, P., Tesch, J., Yang, J.: Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8726–8737 (2023)
7. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 561–578. Springer (2016)
8. Brachmann, E., Wynn, J., Chen, S., Cavallari, T., Monszpart, Á., Turmukhambetov, D., Prisacariu, V.A.: Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. arXiv preprint arXiv:2404.14351 (2024)
9. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on robotics **32**(6), 1309–1332 (2016)
10. Cho, J., Youwang, K., Oh, T.H.: Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In: European Conference on Computer Vision. pp. 342–359. Springer (2022)
11. Choi, H., Moon, G., Chang, J.Y., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1964–1973 (2021)
12. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 769–787. Springer (2020)
13. Choutas, V., Bogo, F., Shen, J., Valentin, J.: Learning to fit morphable models. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI. pp. 160–179. Springer (2022)
14. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. In: International Conference on Machine Learning. pp. 7480–7512. PMLR (2023)

15. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation from multiple views. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7792–7801 (2019)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. arXiv preprint arXiv:2305.20091 (2023)
18. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4318–4329 (2021)
19. He, C., Saito, J., Zachary, J., Rushmeier, H., Zhou, Y.: Nemf: Neural motion fields for kinematic animation. Advances in Neural Information Processing Systems **35**, 4244–4256 (2022)
20. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
21. Henning, D.F., Choi, C., Schaefer, S., Leutenegger, S.: Bodyslam++: Fast and tightly-coupled visual-inertial camera and human motion tracking. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3781–3788. IEEE (2023)
22. Henning, D.F., Laidlow, T., Leutenegger, S.: Bodyslam: joint camera localisation, mapping, and human motion tracking. In: European Conference on Computer Vision. pp. 656–673. Springer (2022)
23. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. arXiv preprint arXiv:2404.15506 (2024)
24. Huang, Y., Bogo, F., Lassner, C., Kanazawa, A., Gehler, P.V., Romero, J., Akhter, I., Black, M.J.: Towards accurate marker-less human shape and pose estimation over time. In: 2017 international conference on 3D vision (3DV). pp. 421–430. IEEE (2017)
25. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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)
26. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5579–5588 (2020)
27. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In: 2021 International Conference on 3D Vision (3DV). pp. 42–52. IEEE (2021)
28. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018)

29. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5614–5623 (2019)
30. Kaufmann, M., Song, J., Guo, C., Shen, K., Jiang, T., Tang, C., Zárate, J.J., Hilliges, O.: Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14632–14643 (2023)
31. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. arXiv preprint arXiv:2312.02145 (2023)
32. Keller, M., Zuffi, S., Black, M.J., Pujades, S.: Osso: Obtaining skeletal shape from outside. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20492–20501 (2022)
33. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
34. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5253–5263 (2020)
35. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11127–11137 (2021)
36. Kocabas, M., Yuan, Y., Molchanov, P., Guo, Y., Black, M.J., Hilliges, O., Kautz, J., Iqbal, U.: Pace: Human and camera motion estimation from in-the-wild videos. arXiv preprint arXiv:2310.13768 (2023)
37. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: 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. pp. 2252–2261 (2019)
38. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4501–4510 (2019)
39. Li, J., Bian, S., Xu, C., Liu, G., Yu, G., Lu, C.: D &d: Learning human dynamics from dynamic camera. In: European Conference on Computer Vision. pp. 479–496. Springer (2022)
40. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3383–3393 (2021)
41. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V. pp. 590–606. Springer (2022)
42. Li, Z., Wallace, E., Shen, S., Lin, K., Keutzer, K., Klein, D., Gonzalez, J.: Train big, then compress: Rethinking model size for efficient training and inference of transformers. In: International Conference on machine learning. pp. 5958–5968. PMLR (2020)
43. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1954–1963 (2021)
44. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12939–12948 (2021)

45. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
46. Liu, M., Yang, D., Zhang, Y., Cui, Z., Rehg, J.M., Tang, S.: 4d human body capture from egocentric video via 3d scene grounding. In: 2021 international conference on 3D vision (3DV). pp. 930–939. IEEE (2021)
47. Loper, M., Mahmood, N., Black, M.J.: Mosh: motion and shape capture from sparse markers. ACM Trans. Graph. **33**(6), 220–1 (2014)
48. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015)
49. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
50. Luo, Z., Golestaneh, S.A., Kitani, K.M.: 3d human motion estimation via motion compression and refinement. In: Proceedings of the Asian Conference on Computer Vision (2020)
51. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
52. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 international conference on 3D vision (3DV). pp. 506–516. IEEE (2017)
53. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 752–768. Springer (2020)
54. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE transactions on robotics **33**(5), 1255–1262 (2017)
55. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV). pp. 484–494. IEEE (2018)
56. Pan, B., Shen, B., Rempe, D., Paschalidou, D., Mo, K., Yang, Y., Guibas, L.J.: Copilot: Human collision prediction and localization from multi-view egocentric videos. arXiv preprint arXiv:2210.01781 (2022)
57. Park, H.S., Shiratori, T., Matthews, I., Sheikh, Y.: 3d trajectory reconstruction under perspective projection. International Journal of Computer Vision **115**, 115–135 (2015)
58. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: 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. pp. 10975–10985 (2019)
59. Pavlakos, G., Malik, J., Kanazawa, A.: Human mesh recovery from multiple shots. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1485–1495 (2022)
60. Pavlakos, G., Weber, E., Tancik, M., Kanazawa, A.: The one where they reconstructed 3d humans and environments in tv shows. In: European Conference on Computer Vision. pp. 732–749. Springer (2022)

61. Peng, X.B., Kanazawa, A., Malik, J., Abbeel, P., Levine, S.: Sfv: Reinforcement learning of physical skills from videos. ACM Transactions On Graphics (TOG) **37**(6), 1–14 (2018)
62. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10985–10995 (2021)
63. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11488–11499 (2021)
64. Rempe, D., Luo, Z., Bin Peng, X., Yuan, Y., Kitani, K., Kreis, K., Fidler, S., Litany, O.: Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13756–13766 (2023)
65. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
66. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. arXiv preprint arXiv:2008.08324 (2020)
67. Saini, N., Huang, C.H.P., Black, M.J., Ahmad, A.: Smartmocap: Joint estimation of human and camera motion using uncalibrated rgb cameras. IEEE Robotics and Automation Letters (2023)
68. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
69. Shen, X., Yang, Z., Wang, X., Ma, J., Zhou, C., Yang, Y.: Global-to-local modeling for video-based 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8887–8896 (2023)
70. Shin, S., Kim, J., Halilaj, E., Black, M.J.: Wham: Reconstructing world-grounded humans with accurate 3d motion. arXiv preprint arXiv:2312.07531 (2023)
71. Sminchisescu, C., Telea, A.C.: Human pose estimation from silhouettes. a consistent approach using distance level sets. In: 10th International Conference on Computer Graphics, Visualization and Computer Vision (WSCG'02). vol. 10 (2002)
72. Smith, C., Charatan, D., Tewari, A., Sitzmann, V.: Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. arXiv preprint arXiv:2404.15259 (2024)
73. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 744–760. Springer (2020)
74. Starke, S., Zhao, Y., Zinno, F., Komura, T.: Neural animation layering for synthesizing martial arts movements. ACM Transactions on Graphics (TOG) **40**(4), 1–16 (2021)
75. Stenum, J., Cherry-Allen, K.M., Pyles, C.O., Reetzke, R.D., Vignos, M.F., Roemmich, R.T.: Applications of pose estimation in human health and performance across the lifespan. Sensors **21**(21), 7315 (2021)
76. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 573–580. IEEE (2012)
77. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11179–11188 (2021)

78. Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M.J.: 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. pp. 8856–8866 (2023)
79. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: 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. pp. 13243–13252 (2022)
80. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Advances in neural information processing systems **34**, 16558–16569 (2021)
81. Ugrinovic, N., Pan, B., Pavlakos, G., Paschalidou, D., Shen, B., Sanchez-Riera, J., Moreno-Noguer, F., Guibas, L.: Multiphys: Multi-person physics-aware 3d motion estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2331–2340 (2024)
82. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
83. Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European conference on computer vision (ECCV). pp. 601–617 (2018)
84. Wan, Z., Li, Z., Tian, M., Liu, J., Yi, S., Li, H.: Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13033–13042 (2021)
85. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7464–7475 (2023)
86. Wang, J., Liu, L., Xu, W., Sarkar, K., Theobalt, C.: Estimating egocentric 3d human pose in global space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11500–11509 (2021)
87. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13031–13040 (2023)
88. Wang, Y., Daniilidis, K.: Refit: Recurrent fitting network for 3d human recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14644–14654 (2023)
89. Wei, W.L., Lin, J.C., Liu, T.L., Liao, H.Y.M.: Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13211–13220 (2022)
90. Weiss, A., Hirshberg, D., Black, M.J.: Home 3d body scans from noisy image and range data. In: 2011 International Conference on Computer Vision. pp. 1951–1958. IEEE (2011)
91. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10974 (2019)

92. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6184–6193 (2020)
93. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems **35**, 38571–38584 (2022)
94. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. arXiv preprint arXiv:2401.10891 (2024)
95. Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling human and camera motion from videos in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21222–21232 (2023)
96. Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11038–11049 (2022)
97. Zanfir, A., Bazavan, E.G., Zanfir, M., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Neural descent for visual 3d human pose and shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14484–14493 (2021)
98. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12104–12113 (2022)
99. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11446–11456 (2021)
100. Zhang, S., Ma, Q., Zhang, Y., Aliakbarian, S., Cosker, D., Tang, S.: Probabilistic human mesh recovery in 3d scenes from egocentric views. arXiv preprint arXiv:2304.06024 (2023)
101. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In: European Conference on Computer Vision. pp. 180–200. Springer (2022)

F Additional Experiments

F.1 Naive Combination: DROID + ZoeDepth

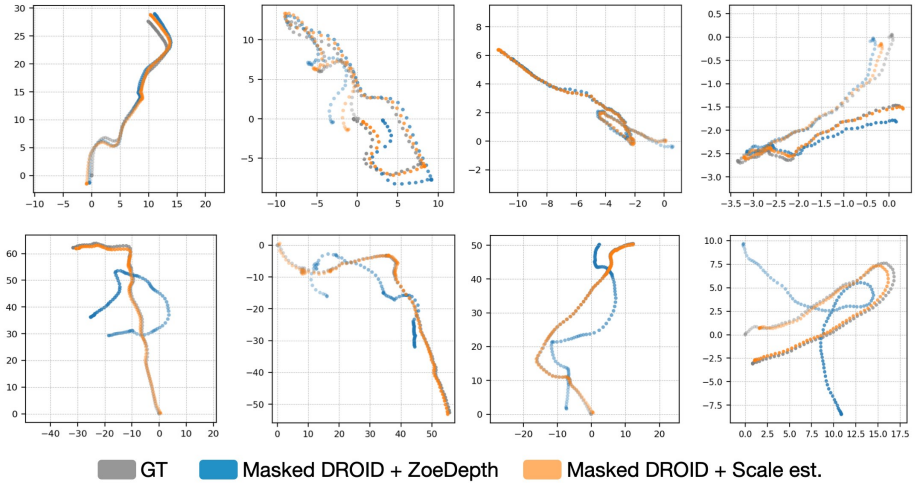

Fig. 7: Camera trajectory. Top row: examples where the naive combination achieves comparable results. Bottom row: naive combination leads to large error.

Our method TRAM derives camera motion scale from the background, by using a robust optimization procedure to align SLAM depth with predicted metric depth (Sec 3.3). As an alternative, we could give metric depth prediction to SLAM along with the input images as pseudo RGB-D inputs. RGB-D SLAM will then return trajectory in metric scale. As indicated in the main text, this naive approach leads to an average ATE-S of $3.09m$, while our method has an average ATE-S of $0.66m$. We visualize this difference in Figure 7. As shown, DROID diverges in roughly half of the sequences due to noisy or spurious depth predictions. Metric depth prediction cannot be treated as RGB-D inputs for a SLAM system.

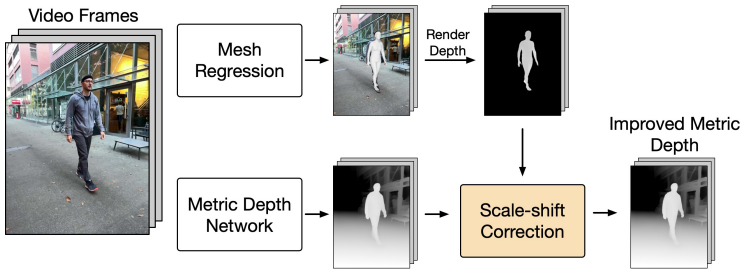

Fig. 8: Improving depth prediction. We render depth from SMPL reconstruction, and use this depth map to correct any shift or scale biases in the metric depth prediction. We scale and shift the depth prediction so that its human region would align with the rendered SMPL depth.

	EMDB 2 (ATE-S)			
Scale Estimation using	Short(5)	Medium(10)	Long(10)	Average
ZoeDepth	0.48	0.62	**0.78**	**0.66**
ZoeDepth + shift correction	0.37	**0.36**	1.41	0.78
ZoeDepth + scale-shift correction	**0.35**	0.37	1.43	0.79

Table 6: Camera scale estimation. Using SMPL depth rendering to correct scale and shift in depth prediction produces mixed results.

F.2 Using SMPL Depth to Improve ZoeDepth

We have shown that ZoeDepth prediction is not always accurate. Particularly, there could be shift and scale biases. In such cases, the metric depth prediction can be regarded as affine-invariant depth prediction. If there are objects of known depth in the image, we can use them to correct the shift and scale. Can we use human mesh reconstruction to help correct the biases? Specifically, could we estimate shift and scale variables s and t to correct depth prediction $\hat{D} = s*D+t$?

Figure 8 illustrates this approach. We solve for the scale and shift correction by aligning the human region in the depth prediction to the rendered depth from SMPL reconstruction, through energy minimization similar to the robust optimization in Sec 3.3 of the main text. We report the quantitative results in Table 6. We observe mixed results: it improves scale estimation in some sequences but decreases accuracy in others. Specifically, we observe that it decreases ATE-S (better) by 20% in 10/25 sequences but increases ATE-S (worse) by 20% in 5/25 sequences. The average ATE-S is slightly worse, because worse cases happen to be long sequences, so a small error in scale estimation could lead to a much higher translation error.

The effectiveness of this approach is also influenced by the accuracy of the mesh reconstruction. If the predicted human shape is more accurate, it will be more effective. Inaccurate shape prediction (e.g., the predicted human being is taller than the ground truth) will produce inaccurate depth rendering.

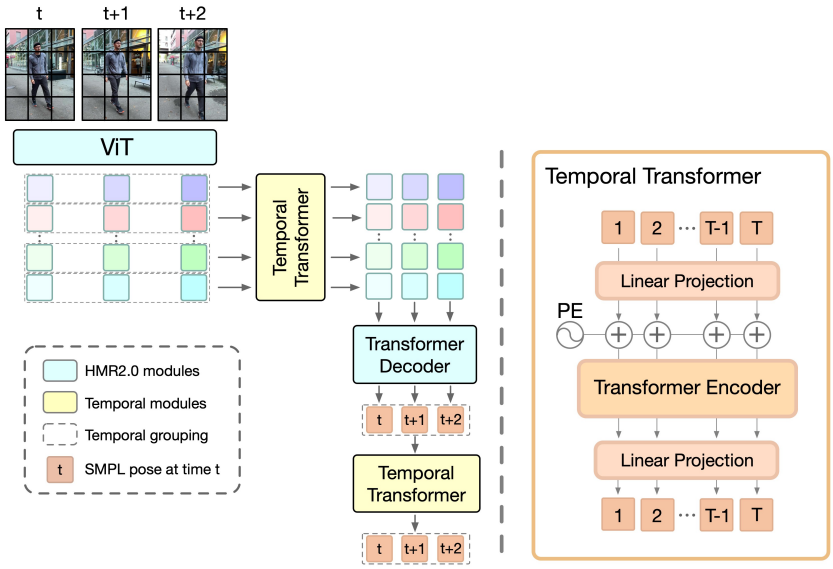

Fig. 9: Architecture of VIMO. Left: the detailed architecture of VIMO, with the yellow blocks denoting the new temporal components. Right: the architecture of the two temporal transformers.

G Implementation Details

G.1 Architecture

We show a more detailed view of the VIMO architecture in Figure 9. VIMO interleaves spatial and temporal modules. Both temporal transformers have 6 layers and 4 multi-head attention. The first temporal transformer (image domain) has an embedding dimension of 512, while the second temporal transformer (motion domain) has an embedding dimension of 384.

G.2 Datasets

We use 3DPW, Human3.6M, and BEDLAM to train our video transformer VIMO. We evaluate on 3DPW and EMDB. **3DPW** is an in-the-wild dataset providing ground truth 3D pose annotations acquired with IMU and videos. 2D and 3D joints are generated from the pose annotation. **Human3.6M** is an indoor multi-view dataset with 2D and 3D joint annotation. Additionally, we use SMPL recovered using MoSH for this dataset. **BEDLAM** is a large synthetic dataset rendered with Unreal Engine 5 and SMPL. Therefore, it has the most accurate SMPL pose and shape. **EMDB** is an in-the-wild dataset with accurate SMPL and trajectory annotations recovered with electromagnetic sensors.

During training, we sample sequences of 16 frames from the three datasets. There are about 1.3k sequences from 3DPW, 19k from Human3.6M, and 305k

from BEDLAM (30fps). So we sample sequences unequally from each dataset to guarantee a good mix of real and synthetic data, with the following ratio: [3DPW: 16.5%, Human3.6M: 16.5%, BEDLAM: 67%].

G.3 ORB-SLAM2

We use the open source ORB-SLAM2 implementation released by the authors in https://github.com/raulmur/ORB_SLAM2. For the masked evaluation, we first process the images in the dataset by setting all pixels within the human masks to a value of 255. We run the entire ORB-SLAM2 pipeline including camera tracking, point reconstruction, and loop closure. We specify our configuration based on the default monocular SLAM parameters for the TUM RGB-D dataset provided in the code and increase the number of features detected at each frame to 4,000. Additionally, because the EMDB videos sometimes demonstrate low contrast with a uniform background, we slightly lowered the minimum fast feature threshold per image patch (more details in the ORB-SLAM2 configuration documentation). Despite these efforts, we still observed tracking failures due to a fast-moving camera as well as textureless background regions. Compared to ORB-SLAM2, DROID performed better in handling these areas because they do not rely on distinctive sparse features; instead, DROID uses optical flow to guide correspondence, which shares the benefits of low-texture texture handling with direct SLAM methods. For loop closure, we use the bag-of-words vocabulary provided in the official repository.

G.4 Training

Acceleration. The training of video models is costly. Because the backbone is often frozen, previous methods pre-compute the features output by the backbone and use them as input to finetune the upper layers (including new components). While this approach reduces forward time, it is impossible to apply data augmentation. To address this issue, we do not pre-compute features but use two other methods: pre-cropping images and half-precision backbone inference. The datasets provide high-resolution images which could take longer to load, crop, and resize. Pre-cropping the images and saving them as crops reduces loading time. Using crops has a different disadvantage: data augmentation such as random rotation and scaling will produce black borders. While it could potentially reduce the effectiveness (it's not clear the extent), it is still better than no augmentation. Secondly, we use half-precision inference for the backbone, which reduces forward time. Since we do not finetune the backbone, using half-precision will not affect the training.

Data Augmentation. We apply standard data augmentation including rotation, scaling, horizontal flipping, color jittering, and occlusion. For video model training, all augmentations except occlusion are applied consistently in the same sequence. For example, the same degree of random rotation should be applied for all 16 frames of a sequence. However, each frame has an independent and equal chance of having occlusion augmentation.