

LAMP: Localization Aware Multi-camera People Tracking in Metric 3D World

Nan Yang Julian Straub Fan Zhang Richard Newcombe Jakob Engel Lingni Ma

Meta Reality Labs Research

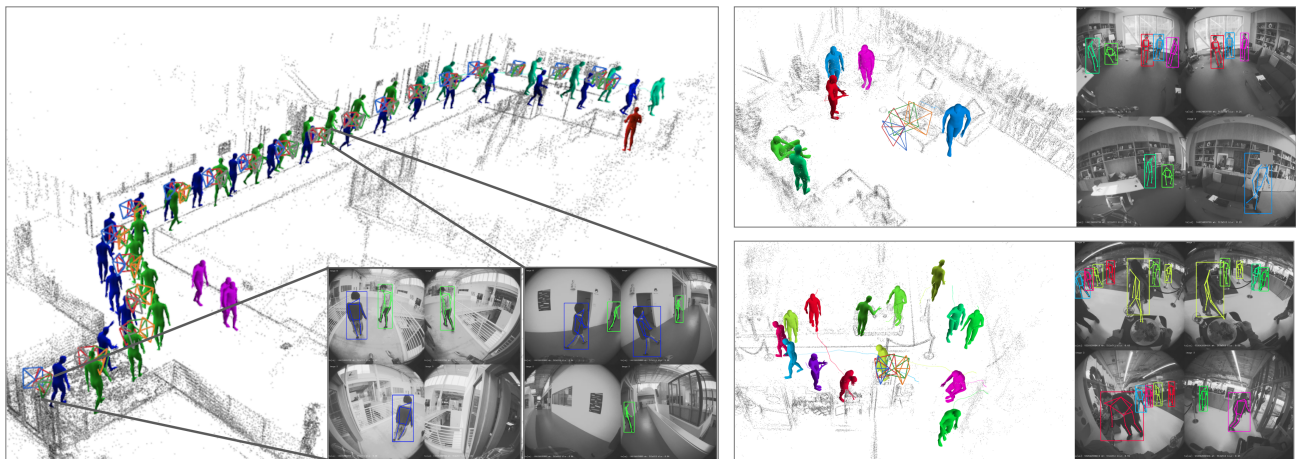


Figure 1. **We propose LAMP, the first method to track human motion with multi-camera headsets in the metric 3D world.** (Left) LAMP persistently track the body motion over a long time, where the 2D observations constantly switching between different cameras across time. (Right) LAMP tracks multiple people simultaneously in real-time across a ultra-wide field-of-view by combining all cameras. Please refer to the *supplementary video* for the real-time demonstrations.

Abstract

Tracking 3D human motion from egocentric multi-camera headset is challenged by severe egomotion, partial visibility or occlusions and lack of training data. Existing methods designed for monocular video often require static or slowly-moving cameras and cannot efficiently leverage multi-view, calibrated and localized input. This makes them brittle and prone to fail on dynamic egocentric captures. We propose **LAMP (Localization Aware Multi-camera People Tracking)**: a novel, simple framework to solve this via early disentanglement of observer and target motion. LAMP introduces a two-step process. First, we leverage the known device 6 DoF motion and calibration to convert detected 2D body keypoints from all cameras over a temporal window into a unified 3D world reference frame. Second, an end-to-end-trained spatio-temporal transformer fits 3D human motion directly to this 3D ray cloud. This “lift-then-fit” approach allows LAMP to learn and leverage a natural human motion prior in the world-space, as well as providing an elegant framework to flexibly incorporate information from

multiple temporally asynchronous, partially observing and moving cameras. LAMP achieves state-of-the-art results on monocular benchmarks, while significantly outperforming baselines for our targeted egocentric setting. Project page: <https://facebookresearch.github.io/LAMP>.

1. Introduction

Augmented reality and smart glasses [1, 15, 49] promise to be persistent, context-aware assistants, enhancing daily life by deeply understanding the user’s environment and acting as effective front-ends that connect a wearer to powerful AI systems. A critical component of such contextual AI systems is social understanding, i.e., the ability to perceive people around the user, what they are doing, and how they are interacting with the user or each other. Understanding 3D human motion is a fundamental component to understanding these higher-level social cues.

Tracking 3D human body motion from videos has been intensively studied over the past decades with impressive progress [20, 23, 67, 68, 78, 79]. However, tracking peo-

ple observed by egocentric modern headsets [1, 15, 50, 51] presents novel challenges that often render existing methods ineffective, due to three major factors. First, headsets are subject to *significant 6-DoF egomotion* from constant, rapid head movement of wearers. This breaks many state-of-the-art motion tracking algorithms, which assume or heavily rely on static or slow-moving camera motion [3, 56, 89]. More fundamentally, these methods attempt to track the motion between observer and target – which can work well when the observer intentionally follows or captures the target (correlated motion), but breaks when these are uncorrelated and subject to independent motion patterns. Second, modern egocentric headsets [1, 15, 49–51] are designed to be *multiview camera rigs*. In order to cover a large field of view, each individual camera captures a different viewing direction, with partial stereo overlapping to a subset of all cameras, as shown in Fig. 3. This means, a person may be fully or partially observed by a single or multiple cameras at each point in time, while the observations also switch from one camera to another across the field-of-view over time. These properties are poorly considered by many existing solutions, which are designed for monocular input. Consequentially, partial single-view observations lead to frequent tracking loss, camera hand-off scenarios are ineffective handled via late fusion, and the results are subject to scale ambiguity inherit to all monocular 3D tracking problems. Third, most existing methods demand a large-scale video data annotated with 3D human motion for training. This type of data is sparse and requires tremendous efforts to collect [25, 26, 88], while synthetic data [5, 84] often lacks realistic moving camera motions. At the same time, modern headsets update camera configurations in every new generations and across manufacturers – making it impractical to built sufficiently large training datasets for a specific device. This is the core reason why most methods that rely on raw video pixels as input to ML models focus on monocular, un-posed and un-calibrated input as lowest common denominator.

To address these challenges, we introduce LAMP: an *early world-space ray lifting* paradigm to track multiple people in metric 3D world with modern egocentric headsets. LAMP takes a posed multi-view video clip as input, and outputs the 3D parameterized body pose [43] per timestamp per person. To achieve this, the method first detects the 2D keypoints [81, 82] independently per person per camera per timestamp. The 2D keypoints are then back-projected and posed by the 6-DoF camera poses into 3D rays, and associated by the targets’ identity across time. The resulting grouped 3D ray cloud is processed by LAMP-Net, a spatial-temporal transformer to output the parameterized body motion for each timestamp. By using the posed 3D rays as input to train our model, LAMP achieves two explicit factorization. First the headset motion is factored out from tracking targets’ motion by using the 6-DoF device localiza-

tion and camera calibrations, which are both available for modern AR or VR headsets by running highly optimized the state-of-the-art visual-inertial odometry (VIO) or full SLAM systems [13, 15, 53, 54] achieving remarkable accuracy and reliability [32]. This factorization enables the entire subsequent fitting problem to fully focus on learning the prior of how people move, and naturally stabilizes and anchors the estimated body motion in the metric 3D world. Second, 2D keypoints detection on raw image pixels is factored out cleanly from 3D motion fitting, thus allowing to leverage existing 2D, image-based human and keypoints detections algorithms [8, 41, 42]. This factorization naturally handles partial observations by individual cameras, and allows for seamless cross-camera hand-off scenarios to derive consistent tracking result. Further, it allows to re-use datasets across different rig-layouts: Since LAMP-Net only requires 2D keypoint observations as input, we can simulate the training data for arbitrary multi-camera headset configurations from arbitrary existing motion datasets [45, 46]. The contribution of our work is summarized as follows.

- We introduce LAMP, a novel system that *lifts and tracks multiple people* directly in the metric 3D world frame from egocentric multi-camera videos by exploiting known 6-DoF camera poses from modern headsets.
- We propose an *early world-space ray lifting* formulation that factors out headset egomotion by lifting 2D keypoints to 3D ray cloud before spatio-temporal reasoning, enabling training from simulated data and naturally supporting multi-view inputs.
- We demonstrate that LAMP significantly outperforms monocular approaches, effectively leveraging multi-view and posed inputs to improve both tracking accuracy and field-of-view coverage. We further show that LAMP *matches* state-of-the-art monocular baselines when reduced to a monocular input rig for a fair comparison.
- We intend to publicly release our model and code to facilitate future research.

2. Related Work

Single image 3D human pose estimation. A large body of work focuses on predicting 3D human pose from a single 2D image [6, 47, 52, 58, 59, 77]. Early works [7, 69] employ a classical optimization approach using a parametric human model, such as SMPL [43] and SMPL-X [60] and minimize the energy with respect to the image observations, such as 2D keypoints and silhouettes. With the advance of deep learning, regression approaches [20, 22, 30, 39, 57] become more popular as they learn human pose and shape priors from large datasets [5, 21, 46], delivering strong generalization capabilities and robustness. However, these single-image models focus on estimating the 3D human pose in the camera coordinate frame, without considering the camera movement or temporal consistencies.

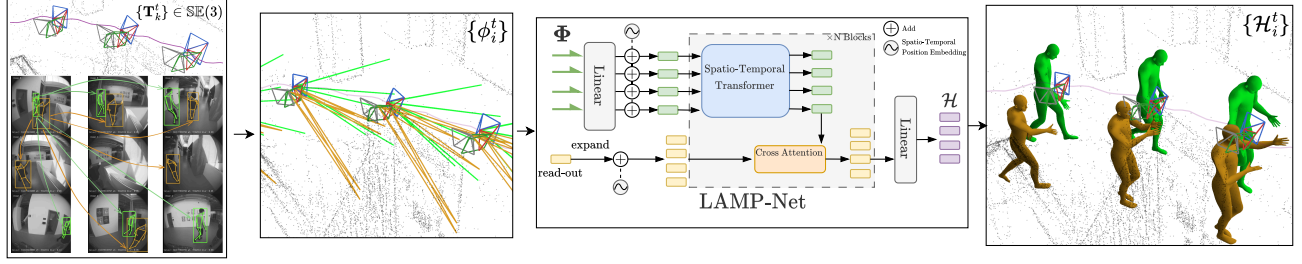


Figure 2. **Method overview.** From egocentric multi-camera video with known 6-DoF poses $\{\mathbf{T}_k^t \in \text{SE}(3)\}$, we detect 2D boxes and keypoints and associate them over time to form per-person tracks as shown in the first subplot. Using the known intrinsics/extrinsics, the 2D keypoints of each track are *lifted* to a sequence of spatio-temporal posed 3D ray clouds in a gravity-aligned reference frame as shown in the second subplot. We then train the LAMP-Net to perform spatio-temporal reasoning which maps ray cloud $\{\phi_i^t\}$ to world-grounded body motion $\{\mathcal{H}_i^t\}$ as shown in the last subplot.

Video-based 3D human pose estimation. Temporal cues are widely exploited to improve stability and reduce per-frame ambiguity in 3D human motion estimates [2, 28, 38, 48, 61, 73, 87, 90, 91]. Most pipelines extract *per-frame* features and then aggregate them over time using temporal convolutions [23, 61, 87], recurrent models [28, 44], or Transformers [38, 66, 76, 93]. While effective for enforcing temporal smoothness, the majority of these methods reason in the *camera* coordinate system and do not explicitly account for camera motion or maintain a consistent world reference, making them ill-suited for scenarios with substantial 6-DoF egomotion (e.g., head-mounted cameras).

Multi-view 3D pose estimation. Many methods reconstruct 3D human pose from multiple, synchronized, and calibrated static cameras, typically via triangulation or volumetric feature fusion [9, 11, 27, 40, 62, 64, 70, 75, 80, 86]. These approaches assume a fixed multi-camera rig set up, and sufficient spatial coverage to maintain visibility, which limits their applicability beyond studio-style setups. In contrast, our scenario uses a *moving* multi-camera rig captured from head mounted devices that casually captures the environment with rapid camera motions.

World-coordinate 3D human pose estimation. Estimating human motion in a fixed world frame is essential for perceiving and interacting with the physical environment. Recent work has begun addressing world-frame recovery from moving cameras. Optimization-based methods such as SLAHMR [83] and PACE [29] refine poses and trajectories but are ill-suited for real-time use. Similarly, MV Lift [37] proposes a diffusion network to synthesize multi-view observations from single view 2D skeleton videos, followed by optimization to fit 3D skeletons, which is not real-time capable. WHAM [68] proposes an end-to-end neural model that leverages gyroscope signals to predict world-frame poses. GVHMR [67] introduces a gravity-view coordinate

to stabilize orientation and reduce drift by predicting in a gravity-aligned frame. To obtain metric scale, TRAM [78], WHAC [84], and PromptHMR [79] couple off-the-shelf SLAM [71, 72] with monocular depth [4], then transform camera-frame predictions into the world frame. GloPro [65] assumes known camera poses but still predicts in camera space before transforming; Ray3D [85] employs a ray-based representation under static-camera rigs. In contrast, we *explicitly* exploit the readily available 6-DoF camera poses on egocentric devices [1, 15, 49] and adopt an **early-lifting** paradigm that lifts the 2D keypoints to 3D rays expressed in a local world frame *before* spatio-temporal reasoning. This avoids late composition as the prior works, enables causal real-time inference, and lets the model learn a unified 3D human motion prior in the world coordinate system.

3. Method

Given multi-view videos $\{I_k^t\}_K^T$ captured by a multi-camera headset with $k \in [1, \dots, K]$ cameras over $t \in [1, \dots, T]$ timestamps, LAMP outputs the metric 3D grounded **body motion tracklets** for each person observed by the video grounded in 3D world coordinates. We assume the camera calibration and 6 DoF camera poses $\mathbf{T}_k^t \in \text{SE}(3)$ are known and accurate. To parameterize 3D human motion, we adopt the commonly used SMPL [43] format. For each tracked person i at time t , the body pose is denoted by $\mathcal{H}_i^t := \{\theta_i^t, \beta_i^t, \omega_i^t, \tau_i^t\}$, where $\theta_i^t \in \mathbb{R}^{69}$ is the SMPL pose parameters that represents the joint angles of 23 body joints, $\beta_i^t \in \mathbb{R}^{10}$ is the SMPL shape parameters with the first 10 PCA coefficients, $\omega_i^t \in \text{SO}(3)$ is the global 3D rotation of the root joint at pelvis, $\tau_i^t \in \mathbb{R}^3$ is the global 3D translation of the root joint. We further denote the clip of human motion by $\mathcal{H}_i^T := \{\mathcal{H}_i^t\}_{t=1}^T$. With the function $f_{\text{SMPL}}(\theta, \beta)$ that converts SMPL shape and pose parameters to $\mathcal{J} \in \mathbb{R}^{23}$ joints and $\mathcal{V} \in \mathbb{R}^{6890}$ mesh vertices in SMPL local coordinates, the joints and vertices in the world coordinates for person i

at time t are then computed by

$$\mathcal{J}_i^t, \mathcal{V}_i^t = \exp(\omega_i^t) f_{\text{SMPL}}(\theta_i^t, \beta_i^t) + \tau_i^t. \quad (1)$$

3.1. Overview

Figure 2 illustrate how LAMP employs an *early world-space ray lifting* paradigm to track multiple people over time for the multi-camera headset. Start from individual images, the method first detects 2D bounding boxes of each observed person and estimate the set of 2D keypoints $\{\mathbf{p}_j^t \in \mathbb{R}^2\}$ per bounding box detection. The 2D detections are spatio-temporally associated across K cameras over T timestamps by the identity of tracking targets. After associations, the 2D keypoints are back-projected into 3D and transformed by the known camera poses $\mathbf{T}_k^t$ and camera calibration parameters to obtain a sequence of spatio-temporal ray clouds aligned in the 3D space. For each group of ray clouds, we use LAMP-Net to map the sequence of 3D rays into 3D grounded human motion $\{\mathcal{H}_i^t\}$.

In this formulation, we made two explicit factorizations. First, the 6 DoF camera poses are disentangled from the movements of tracked people. We argue this is a crucial ingredient which enables LAMP-Net to fully focus on learning the motion prior of how people move. Unlike many previous works that are designed to learn or refine a coarse camera motion together with human motion [78, 79], we believe learning camera motion not only make the problem harder to solve, but also is unnecessary for modern headsets, which device tracking is largely considered a solved problem [10, 12, 15, 17, 32, 35, 53, 55]. The second factorization is to separate 2D human detections from the motion learning and directly learn to solve the inverse problem “find the parameterized 3D human motion that is most consistent with a given spatio-temporal ray cloud”. The formulation encourages LAMP-Net to flexibly combine and leverage temporally asynchronous observations for “3D triangulation” by connecting them through the local dynamics imposed by human motion prior. By doing so, LAMP is able to naturally handle partial 2D observations and camera hand-off scenarios to persist tracking across time and camera views. As a by-product of this factorization, LAMP can simply be re-trained for a new camera configuration by simulating 2D observations for the targeted camera layout from arbitrary motion datasets. This solves a key challenge in data scarcity for posed multi-view devices, which is the core reason why most research focuses on the monocular setting. In the remaining section, we further describe the ray association, 2D-to-3D multi-view lifting and LAMP-Net training for spatio-temporal ray fitting in more details.

3.2. Tracklet spatio-temporal association

For a timestep t , we obtained the 2D bounding boxes of all visible people using 2D detection algorithms [18, 63, 81].

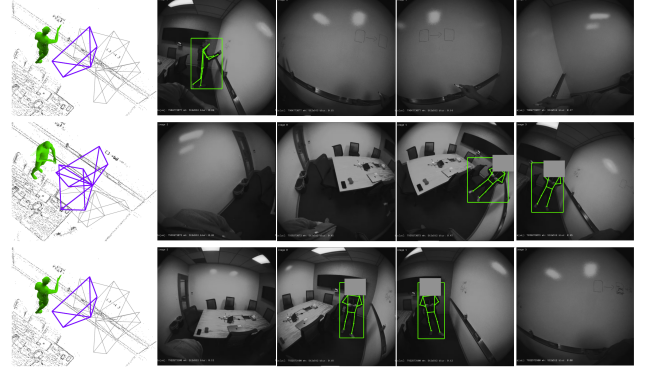


Figure 3. **Multi camera tracking.** LAMP seamlessly estimates body motion for a person across several “camera-handoffs” for a sequence captured from an Project Aria Gen2 glasses and using all four available monochrome cameras. Our formulation allows to seamlessly combine all available observations in a single model inference call to fitting a full 4s 3D motion snippet.

For association, we solve a *bipartite matching problem* between these 2D bounding boxes and active tracklets from the previous timesteps. This is done by projecting 3D points into the images to obtain an expected observation location and compute the matching cost. Note that this automatically compensates for the headset motion, as tracklets are represented in the world coordinates. Matching observations is then solved with the Hungarian algorithms [3, 34, 89]. When a 2D bounding box fails to associate with others, we instantiate a new tracklet. For real world inference, a tracklet will be deactivated if the person is not observed for more than a threshold of time.

3.3. World aligned ray lifting

For each 2D bounding box, we run ViTPose [82] to detect the MSCOCO [42] 2D keypoints on the cropped images. We then assemble the 3D ray cloud ϕ_i^t by first unprojecting the 2D keypoints into 3D, transforming them by the corresponding camera pose, and normalizing the reference coordinates by the first frame of the temporal window to prepare the input for LAMP-Net inference. For coordinates normalization, we define a local coordinates L_t based on the pose of the camera $i = 0$, which effectively aligns the reference with gravity as estimated by visual inertial odometry (VIO)

$$\mathbf{T}_{W \leftarrow L_T}^t := \text{gravity_align}(\mathbf{T}_{W \leftarrow C_0}^t) \in \mathbb{SE}(3). \quad (2)$$

Together, the compute of 3D rays into the local coordinate system follows

$${}^c\phi_j^t := \mathbf{T}_{L_T \leftarrow W} \cdot \mathbf{T}_{W \leftarrow C_k}^t \cdot \pi^{-1}(\mathbf{p}_j^t), \quad (3)$$

where $\pi^{-1}: \Omega \rightarrow \mathbb{R}^3$ denotes the calibrated camera unprojection function of a pixel to a unit ray, and $\mathbf{T}_{W \leftarrow C_k}^t \in$

$\text{SE}(3)$ denotes the known camera pose of camera k at time t . Combined with the respectively transformed camera centers

$${}^o\phi_t^k := \mathbf{T}_{L_T \leftarrow W} \cdot \mathbf{T}_{W \leftarrow C_k}^t \cdot [0, 0, 0, 1]^\top, \quad (4)$$

this results in one 3D ray per 2D observation. Finally, we stack all rays within the window, parametrized as 6-dimensional Plücker rays concatenated with the confidence score from the 2D keypoint detector. The final result is a tensor $\Phi \in \mathbb{R}^{T \times K \times J \times 7}$, where $J = 17$ is the number of MSCOCO human keypoints and K is the number of cameras. The values corresponds to the missing 2D keypoints are set to zero for training and inference with LAMP-Net, described in the next section.

3.4. Spatio-temporal ray fitting with LAMP-Net

The function to map spatio-temporal 3D ray cloud to the corresponding 3D human motion is modeled by LAMP-Net. To this end, we use a spatio-temporal transformer architecture [74, 93]. The transformer takes the sequential ray cloud as the input, and regresses the SMPL motion parameters [43] $\{\mathcal{H}_i^t\}$ for each timestamp. The spatio-temporal encoder performs self-attention across both spatial (i.e., by joint) and temporal (i.e., by frame) dimensions to jointly estimate body shape and motion dynamics. A learnable read-out embedding is expanded and added with the time encoding to form the query tokens in a cross-attention decoder. Unlike previous works where network readouts interact only with the last encoder layer [19], our decoder performs cross-attention at each encoder block, allowing the readout embeddings to iteratively aggregate motion and geometry information across different feature hierarchies. We found that this multi-level interaction improves temporal stability and convergence. For training, we adopt the 6D rotation [92] to parameterize 3D rotation. Note that since we normalize the input by the local coordinates L_T , the predicted motions needs to be transformed back to the original 3D world reference.

Training losses. Inspired by prior work [20, 78], the network training losses penalize the deviation to the ground-truth SMPL parameters $\mathcal{H}_i^t$, the 3D joint positions $\mathcal{J}$, 3D vertex positions $\mathcal{V}$, as well as the joint velocity $\mathcal{D}_J$. The loss function is defined as

$$\mathcal{L} = \lambda_{\text{SMPL}} \mathcal{L}_{\text{SMPL}} + \lambda_{3\text{D}} \mathcal{L}_{3\text{D}} + \lambda_{\text{V}} \mathcal{L}_{\text{V}} + \lambda_{\text{vel}} \mathcal{L}_{\text{vel}}, \quad (5)$$

with

$$\begin{aligned} \mathcal{L}_{\text{SMPL}} &= \frac{1}{T} \sum_{t=1}^T \|\hat{\mathcal{H}}^t - \mathcal{H}^t\|_2^2 \\ \mathcal{L}_{3\text{D}} &= \frac{1}{T} \sum_{t=1}^T \|\hat{\mathcal{J}}_t - \mathcal{J}_t\|_F^2 \\ \mathcal{L}_{\text{V}} &= \frac{1}{T} \sum_{t=1}^T \|\hat{\mathcal{V}}_t - \mathcal{V}_t\|_F^2 \\ \mathcal{L}_{\text{vel}} &= \frac{1}{T-1} \sum_{t=2}^T \|\hat{\mathcal{D}}_t - \mathcal{D}_t\|_F^2. \end{aligned}$$

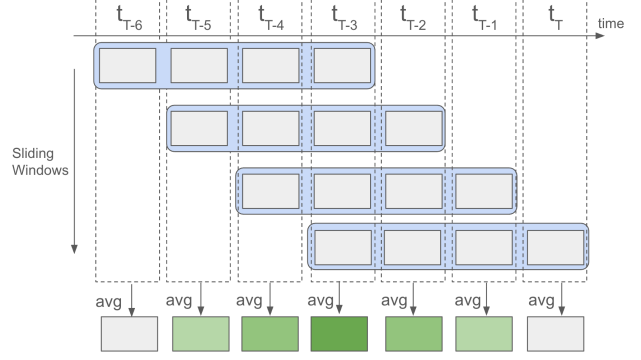


Figure 4. **Sliding window inference.** We propose a simple and light-weight sliding-window inference strategy by averaging the same-time-stamp pose predictions to get more accurate and stable human motions.

We empirically found that the vertices loss $\mathcal{L}_{\text{V}}$ improves the results even though it is ill-posed for LAMP to recover accurate mesh due to the sparse 2D keypoint observations.

3.5. Non-causal inference with temporal smoothing

LAMP-Net consumes a window of video clips as input, each frame will be processed T times during a strictly causal on-line inference, where the input window increments forward 1 frame at a time to yield the estimate for the latest frame. This inference paradigm is illustrated in Fig. 4. With this nature of inference, smoothing the past estimations by averaging the predictions over time comes with no additional cost network inference per se. Assume the network predictions contain noise, averaging multiple estimates over time is expected to improve the accuracy and reduce jitter. However, smoothing results over sliding window increases the latency of output. At maximum, it can delay the output by $T - 1$ timestamps. In practice, this property of LAMP inference provides a run-time variable that can be used control the tradeoff between accuracy versus latency. In our experiments, we evaluate the impact of smoothing with maximum latency, where the pose at time t is computed by averaging all T variants of predictions available over time.

4. Experiments

4.1. Implementation details

Model. LAMP-Net contains three transformer encoder-decoder blocks with 512 inner dimension. The input is set to be a 4-second temporal window. With video data at 30 Hz, this amounts to in $W = 120$ frames input. The loss weights are set to be $\lambda_{\text{SMPL}} = 1.0$, $\lambda_{3\text{D}} = 5.0$, $\lambda_{\text{V}} = 1.0$, $\lambda_{\text{vel}} = 20.0$. The network is trained for 200 epochs on 4 nodes of NVidia H100 GPUs. Training takes approximately 19 hours to converge. The inference and evaluation run on a single RTX4090 GPU for real-time inference.

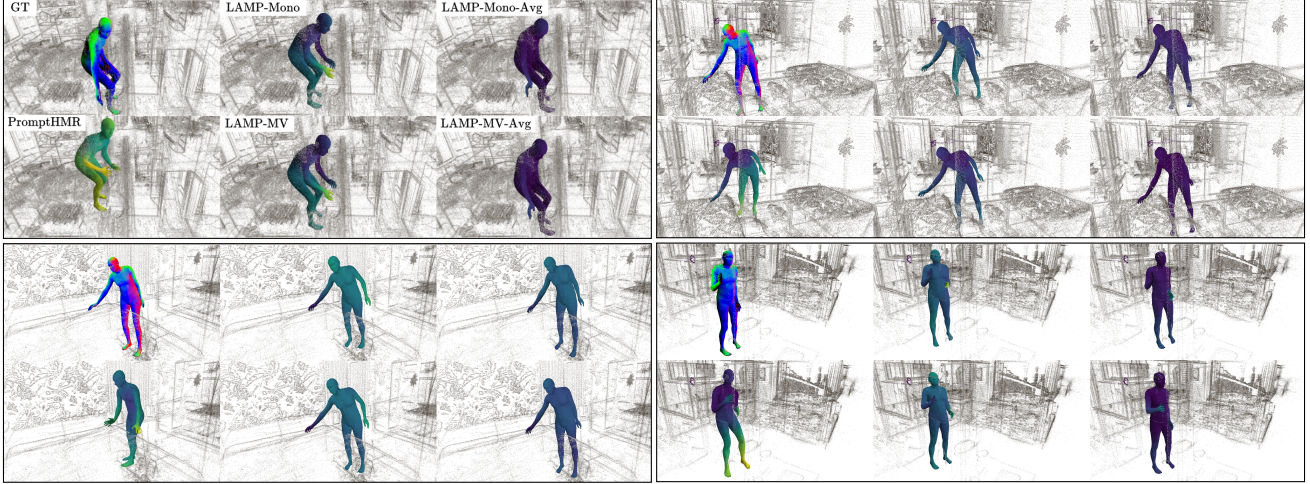


Figure 5. **Qualitative comparison of 3D human motion estimation.** We compare the output PromptHMR [79] against monocular LAMP and multi-camera (MV) LAMP on Nymeria [45] with and without temporal smoothing. Per output SMPL mesh, the vertex colors encode the Euclidean distance to the corresponding ground-truth vertex, with the higher error showed in yellow and lower error in dark purple.



Figure 6. **Tracking coverage versus number of cameras.** (Left): Distribution of the proportion of people tracked per timestamp against the proportion of time. Using more cameras clearly shifts mass toward 1.0, meaning all people are tracked more often. In a dynamic social interaction with three other people, average coverage is 47% for 1-cam, 65% for 2-cam, and 81% for 4-cam. Right: qualitative examples showing how using more cameras improves the tracking coverage.

Dataset. LAMP-Net is trained with Nymeria dataset [45]. The dataset contains 1100 recordings, more than 300 hours of in-the-wild human motion from 264 participants, using multiple Project Aria glasses [15]. The dataset provides ground-truth body motions for subjects using XSens mocap suit, with retargetted SMPL [43] format and we use the observers’ Aria glasses to generate training data. The dataset is then split into 770, 165 and 26 sequences for training, validation and testing. The test set covers one of each 20 activity scenarios to evaluate diverse motion. For training data preparation, we project the 3D ground-truth body joints onto the observer cameras using the ground-truth intrinsics and extrinsics. We apply extensive data augmentation to simulate real world noisy 2D detections and occlusions. We also augment the headset motion to further increase the diversity. The data augmentation is further detailed in the supplemental. In addition, we also evaluate LAMP with EMDB [24], a widely-adopted dataset in benchmarking 3D human tracking. Since EMDB is a small dataset with total 40 minutes recordings, we do not train or finetune our model on EMDB. Therefore, our results on EMDB reflect the zero-shot generalization of LAMP.

Simulate arbitrary multi-camera layout. A key benefit of the 3D ray lifting formulation is that LAMP-Net does not require direct pixels or image features for training and inference. Therefore, we can easily simulate 2D keypoint observations by projecting ground-truth 3D joints into virtual cameras. This design enables training with arbitrary 3D human motion dataset without video data [45, 46]. Importantly, this training scheme enables large-scale data synthesis [68] with diverse observer-camera configurations. In our experiments, we qualitatively evaluate this feature of LAMP by simulating a large Aria Gen2 [49] dataset from the Nymeria dataset which is collected with Aria Gen1 [14], and show that the model is able to handle real-world Aria Gen2 dataset, despite never trained with any real Aria Gen2 data. Note that there is a significant update on camera configuration from Aria Gen1 to Aria Gen2.

Metrics. We evaluate LAMP using a range of common metrics, mostly derived from the Mean Per-Joint Position Error (MPJPE) error and reported in millimeters (mm). Here we provide a brief description of different MPJPE variants and propose a new variant W-MPJPE for evaluating motion

Table 1. **Quantitative comparison of LAMP against state-of-the-art algorithms.** We evaluate LAMP on EMDB [24] and Nymeria [45] datasets, where the best results for each dataset is marked in bold, respectively. We refer to GEM [36] for the evaluation results of WHAM [68] and GVHMR [67] on EMDB [24] with GT camera poses. The metrics are reported in millimeter (mm) for all MPJPE variants and foot skating (FS), in percentage (%) for RTE and in $10m/s^3$.

Method	Dataset	MPJPE	PA-MPJPE	WA-MPJPE ₁₀₀	W-MPJPE	RTE	Jitter	FS
WHAM [68]	EMDB	81.6	52.0	131.1	-	4.1	21.0	4.4
GVHMR [67]	EMDB	74.2	44.5	109.1	-	1.9	16.5	3.5
GEM [36]	EMDB	73.0	42.5	69.5	-	0.9	17.7	8.6
PromptHMR [79]	EMDB	68.1	40.1	63.9	278.1	0.4	16.3	3.5
LAMP-mono (ours)	EMDB	82.3	46.3	77.8	165.1	0.2	4.6	3.2
PromptHMR [79]	Nymeria	109.2	66.0	101.6	246.0	0.11	114.1	7.7
LAMP-mono (ours)	Nymeria	92.3	55.5	80.4	203.4	0.09	23.8	3.2
LAMP-mv (ours)	Nymeria	54.8	37.3	58.7	113.3	0.05	21.8	3.6

tracking in metric 3D world. MPJPE measures the average joint position error after removing any translational misalignment of the root joint (pelvis) w.r.t. ground truth. Similarly, PA-MPJPE measures the average joint position error after Procrustes alignment with the ground truth. Both metrics are designed to measure the local similarity of body poses and are computed on a per-frame basis. WA-MPJPE₁₀₀ on the other hand, is designed to infer the global body pose accuracy. Due to the common setting with monocular video input, the SIM(3) alignment is first performed over a 100-frame temporal window before computing the error in order to reduce the impact of scale ambiguity. For multi-camera headset input, we argue that algorithms should be able to resolve the scale ambiguity and algorithms should be directly evaluated against ground truth without any alignment. To this end, we introduce a new metric W-MPJPE, which measures the MPJPE without any alignment. Following previous works [29, 68, 79, 83], we also report jitter, foot skating (FS), and root trajectory error (RTE).

4.2. Comparison to state-of-the-art

We conduct evaluations against recent state-of-the-art methods [36, 67, 68, 79]. The quantitative comparison on EMDB and Nymeria is shown in Tab. 1 and the qualitative comparison on Nymeria is presented in Fig. 5.

Evaluation on EMDB. We first evaluate LAMP on the EMDB-2 test split. EMDB is captured from slowly moving cameras closely following the participant. We use ground-truth camera extrinsics for all the methods and run LAMP with *single-view input* for direct comparison to other methods. Overall, LAMP achieves highest RTE, W-MPJPE and Jitter by a significant margin, indicating superior world-space localization accuracy. Figure 7 also shows the trajectory error overtime comparing LAMP and PromptHMR. In turn, LAMP falls notably behind on MPJPE, PA-MPJPE and WA-MPJPE₁₀₀. We believe this can be attributed to

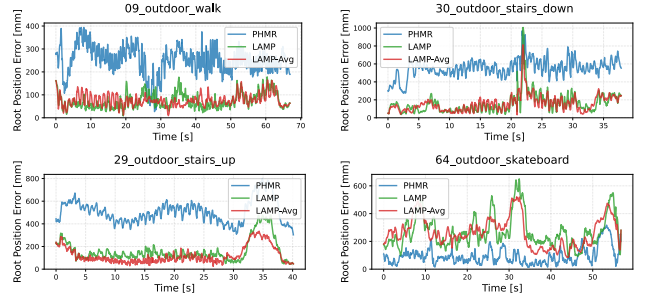


Figure 7. **Root Trajectory Errors on EMDB.** LAMP outperforms PromptHMR on absolute root trajectory accuracy on most sequences. However, on *64_outdoor_skateboard*, LAMP shows inferior result which is likely due to the lack of skateboarding activity in the training data.

our design choice of collapsing raw images into 3D rays to facilitate multi-view aggregation and continuous tracking over time, capabilities which are not captured by these metrics and on this dataset. Please refer to our supplementary materials for the qualitative results on EMDB.

Evaluation on Nymeria. We compare LAMP with PromptHMR [79] on the Nymeria dataset. Nymeria differs from EMDB in 2 key aspects: 1) it contains longer sequences (15 mins versus 1 min per sequence) allowing better evaluation over long-term accuracy; 2) the observer camera contains natural egocentric head motions with frequent rotations. For comparison we adapt PromptHMR to take ground-truth camera poses, rectify, and process the sequences in 1200 frame-chunks. We observe that LAMP notably outperforms PromptHMR on all metrics even in the monocular configuration, where multiview input gives additional boost to accuracy as shown in Fig. 5. Note that for the multi-camera setup we use the keypoint detections from 1 RGB and 2 SLAM cameras from Project Aria glasses (Gen1). This is further discussed in Sec. 4.3. More qualitative results are presented in the supplementary.

Table 2. **Ablation study.** The design choices (first four columns) mean the following. By *posed*, we transform 3D rays by camera poses; by *smooth*, per-frame estimation is averaged over a sliding window; by *simulated*, we simulate 2D keypoint detections from ground truth; and by *multiview*, we use four-camera [49] observations instead of monocular video. Experiments is done with the Nymeria [45] datasets. All MPJPE variants and foot skating (FS) are in millimeter (mm), RTE in percentage (%) and jitter in $10m/s^3$.

variant	posed	smooth	simulate	multiview	MPJPE	PA-MPJPE	WA-MPJPE ₁₀₀	W-MPJPE	RTE	Jitter	FS
var_0					98.5	57.5	95.8	296.3	0.50	93.1	6.1
var_1	✓				98.3	57.6	89.3	209.6	0.09	91.7	5.5
var_2	✓	✓			92.3	55.5	80.4	203.4	0.09	23.8	3.2
var_3	✓	✓	✓		60.4	39.2	71.7	199.8	0.08	21.4	3.5
var_4	✓	✓		✓	54.8	37.3	58.7	113.3	0.05	21.8	3.6
var_5	✓	✓	✓	✓	52.0	34.8	58.2	111.5	0.05	21.4	3.5

Qualitative evaluation on Aria Gen2. To show that our training method enables generalization to different camera layouts, we recorded test sequences with the Project Aria Gen2 headset [49], which mounts 4 CV cameras: a forward-facing stereo pair with substantial overlap plus two side-view cameras, yielding a $\sim 270^\circ$ field of view. LAMP exploits the known intrinsics/extrinsics to fuse all views; as illustrated in Fig. 1, it persistently tracks the people over long trajectories (*Left*) and tracks multiple people in real-time across the wide FoV (*Right*). To quantify the effect of camera count on coverage, we analyze four 5-minute sequences from the Aria Gen2 Pilot dataset [31], comparing 1-, 2-, and 4-camera settings. Figure 6 shows that adding views substantially increases per-timestamp coverage in dynamic social interactions, confirming the benefit of the multi-camera rig.

4.3. Ablation studies

We conduct ablation studies on the Nymeria dataset. We preprocess the dataset to evaluate only frames with sufficient visibility of the tracked person – we observe that the dataset contains complex activities and interactions, and thus – in contrast to EMDb – the tracked person is not always visible and can be temporarily occluded by e.g. walls or furniture. We achieve this by running an off-the-shelf 2D bounding box detector [81] for people on the full video, and filter out any frames in which no person is detected with a 2D bounding-box IoU of at least 0.4 compared to surrounding boxes of the projected 3D joints.

Table 2 reports the effect of each design choice for LAMP. The first row (var_0) is the result of LAMP when *not using known camera motion*, and instead lifting rays into the moving observer frame of reference. Adding LAMP’s 3D Ray-Lifting (var_1) significantly improves all Global Metrics, as well as RTE and foot-skating. This is expected, and indeed a core benefit of our method which allows to fully leverage known observer-motion for accurate 3D localization of the observed person. Averaging overlapping windows further improves all metrics (var_2). To evaluate the impact of multiple cameras, we use Aria Gen1 multi-view camera rig with 2D keypoints detected by ViTPose [82] (var_4). In

addition, we include results with simulated 2D keypoints for both monocular and multi-view experiments (var_3 and var_5). The comparison between var_2 and var_4 shows the improvement by using all 3 cameras vs. when using only a monocular view. Again, using multi-view input yields significant improvement across all metrics except jitter and foot-sliding. An additional observation is that the sim-to-real delta of multi-view inputs (i.e., between var_4 and var_5 is significantly smaller than that of using monocular inputs (i.e., between var_2 and var_3), which suggests, although LAMP-Net is trained with simulation only, using multi-view inputs together with our extensive data augmentation closes the sim-to-real gap notably. Please refer to our supplementary materials for additional experiments on camera pose and 2D keypoints sensitivity, runtime, and tracking performance.

5. Conclusion

LAMP introduces a new approach to tracking people and their body-motion in 3D metric world space. By accepting known 6DoF motion and calibration as readily-available input modality, LAMP provides a simple and flexible solution to disentangling observer- and target-motion, as well as to fuse observations from multi-view video input. By further abstracting raw pixels into 2D rays early-on, LAMP can easily be re-trained for arbitrary rig-layouts facilitating cross-device usage. We demonstrate LAMP’s superior performance in the targeted scenario of in-the-wild egocentric observations of surrounding people – as supposed to intentional/staged capture from static or hand-held cameras.

Limitations LAMP depends on accurate and reliable 6 DoF tracking as input modality, and requires multi-view input to unfold its full potential. While this is commonly available for modern egocentric devices, it does prevent LAMP from being used effectively for monocular mobile phone captures or common internet-sourced recordings. Besides, including more comprehensive pixel-derived information will potentially improve the local human pose accuracy.

Acknowledgements

We would like to thank Federica Bogo, Bharat Bhatnagar, Jinlong Yang, Yuanlu Xu, David Cruso, and Daniel DeTone for their valuable support and fruitful discussions throughout the course of this project.

References

- [1] Apple Inc. Apple Vision Pro. <https://www.apple.com/apple-vision-pro/specs/>. 1, 2, 3
- [2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3395–3404, 2019. 3
- [3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. Ieee, 2016. 2, 4
- [4] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3
- [5] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 2
- [6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 2
- [7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 2
- [8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229, 2020. 2
- [9] He Chen, Pengfei Guo, Pengfei Li, Gim Hee Lee, and Gregory Chirikjian. Multi-person 3d pose estimation in crowded scenes based on multi-view geometry. In *European Conference on Computer Vision*, pages 541–557. Springer, 2020. 3
- [10] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE TPAMI*, 2007. 4
- [11] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3D pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7792–7801, 2019. 3
- [12] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *ECCV*, 2014. 4
- [13] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625, 2018. 2
- [14] Jakob Engel, Kiran Somasundaram, Michael Giese, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brigid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Gajipalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eickenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreeves, Xiqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project Aria: A New Tool for Egocentric Multi-Modal AI Research, 2023. 6
- [15] Jakob Engel, Kiran Somasundaram, Michael Giese, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brigid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 1, 2, 3, 4, 6
- [16] Aaron Ferguson, Ahmed A. A. Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, Igor Santesteban, Javier Romero, Jenna Zarate, Jeongseok Lee, Jinhyung Park, Jinlong Yang, John Doublestein, Kishore Venkateshan, Kris Kitani, Ladislav Kavan, Marco Dal Farra, Matthew Hu, Matthew Cioffi, Michael Fabris, Michael Ranieri, Mohammad Modar-

- res, Petr Kadlecek, Rawal Khirodkar, Rinat Abdrashitov, Romain Prévost, Roman Rajbhandari, Ronald Mallet, Russel Pearsall, Sandy Kao, Sanjeev Kumar, Scott Parrish, Shoou-I Yu, Shunsuke Saito, Takaaki Shiratori, Te-Li Wang, Tony Tung, Yichen Xu, Yuan Dong, Yuhua Chen, Yuanlu Xu, Yuting Ye, and Zhongshi Jiang. Mhr: Momentum human rig, 2025. [1](#)
- [17] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: Fast Semi-Direct Monocular Visual Odometry. 2014. [4](#)
- [18] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. [4](#)
- [19] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. [5](#)
- [20] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Reconstructing and tracking humans with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. [1](#), [2](#), [5](#)
- [21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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. [2](#)
- [22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. [2](#)
- [23] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. [1](#), [3](#)
- [24] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. 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*, pages 14632–14643, 2023. [6](#), [7](#), [1](#)
- [25] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark. *arXiv preprint arXiv:2305.16487*, 2023. [2](#)
- [26] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4d: A video dataset for in-the-wild close human interactions. *Advances in Neural Information Processing Systems*, 37: 107270–107285, 2024. [2](#)
- [27] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1077–1086, 2019. [3](#)
- [28] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. [3](#)
- [29] 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 *International Conference on 3D Vision*, pages 397–408, 2024. [3](#), [7](#)
- [30] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 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*, pages 2252–2261, 2019. [2](#)
- [31] Chen Kong, James Fort, Aria Kang, Jonathan Wittmer, Simon Green, Tianwei Shen, Yipu Zhao, Cheng Peng, Gustavo Solaira, Andrew Berkovich, et al. Aria gen 2 pilot dataset. *arXiv preprint arXiv:2510.16134*, 2025. [8](#)
- [32] Anusha Krishnan, Shaohui Liu, Paul-Edouard Sarlin, Oscar Gentilhomme, David Caruso, Maurizio Monge, Richard Newcombe, Jakob Engel, and Marc Pollefeys. Benchmarking Egocentric Visual-Inertial SLAM at City Scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. [2](#), [4](#)
- [33] Anusha Krishnan, Shaohui Liu, Paul-Edouard Sarlin, Oscar Gentilhomme, David Caruso, Maurizio Monge, Richard Newcombe, Jakob Engel, and Marc Pollefeys. Benchmarking egocentric visual-inertial slam at city scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. [1](#)
- [34] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2: 83–97, 1955. [4](#)
- [35] Stefan Leutenegger. OKVIS2: Realtime Scalable Visual-Inertial SLAM with Loop Closure. *arXiv:2202.09199*, 2022. [4](#), [1](#)
- [36] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. *arXiv preprint arXiv:2505.01425*, 2025. [7](#)
- [37] Jiaman Li, C Karen Liu, and Jiajun Wu. Lifting motion to the 3d world via 2d diffusion. In *Proceedings of the*

- Computer Vision and Pattern Recognition Conference*, pages 17518–17528, 2025. 3
- [38] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13147–13156, 2022. 3
- [39] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 2
- [40] Jiahao Lin and Gim Hee Lee. Multi-view multi-person 3d pose estimation with plane sweep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11886–11895, 2021. 3
- [41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2
- [42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 4
- [43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6): 1–16, 2015. 2, 3, 5, 6, 1
- [44] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3D human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3
- [45] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Gu-zov, Yifeng Jiang, Rowan Postyni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pages 445–465. Springer, 2024. 2, 6, 7, 8, 1
- [46] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. 2, 6
- [47] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 2
- [48] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017. 3
- [49] Meta. Aria Gen 2: An Advanced Research Device for Egocentric AI Research. <https://www.projectaria.com/glasses/>. 1, 2, 3, 6, 8
- [50] Meta Platforms, Inc. Meta Quest 3: Next-Gen Mixed Reality Headset. <https://www.meta.com/quest/quest-3/>. 2
- [51] Microsoft Corporation. HoloLens 2. <https://learn.microsoft.com/en-us/hololens/>. 2
- [52] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2823–2832, 2017. 2
- [53] Anastasios I. Mourikis and Stergios I. Roumeliotis. A multi-state constraint Kalman filter for vision-aided inertial navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3565–3572, Rome, Italy, 2007. 2, 4
- [54] Raul Mur-Artal and Juan D Tardós. ORB-SLAM: An open-source slam system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 2
- [55] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 4
- [56] Alejandro Newell, Peiyun Hu, Lahav Lipson, Stephan R. Richter, and Vladlen Koltun. Comotion: Concurrent multi-person 3d motion. In *International Conference on Learning Representations*, 2025. 2
- [57] Priyanka Patel and Michael J Black. Camerahr: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pages 1562–1571. IEEE, 2025. 2
- [58] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017. 2
- [59] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 2

- [60] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 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*, pages 10975–10985, 2019. 2
- [61] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019. 3
- [62] Zekun Qian, Wei Feng, Feifan Wang, and Ruize Han. Bevtrack: Multi-view multi-human registration and tracking in the bird’s eye view. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3
- [63] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016. 4
- [64] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6040–6049, 2020. 3
- [65] Simon Schaefer, Dorian Henning, and Stefan Leutenegger. Glopro: Globally-consistent uncertainty-aware 3d human pose estimation tracking in the wild. *IROS*, 2023. 3
- [66] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. 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*, pages 8887–8896, 2023. 3
- [67] 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. 1, 3, 7
- [68] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. *arXiv preprint arXiv:2312.07531*, 2023. 1, 3, 6, 7
- [69] Cristian Sminchisescu and Alexandru Telea. Human pose estimation from silhouettes. a consistent approach using distance level sets. In *EPRINTS-BOOK-TITLE*. University of Groningen, Johann Bernoulli Institute for Mathematics and . . . , 2002. 2
- [70] Vinkle Srivastav, Keqi Chen, and Nicolas Padoy. Self-pose3d: self-supervised multi-person multi-view 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2502–2512, 2024. 3
- [71] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and RGB-D cameras. *Advances in Neural Information Processing Systems*, 34:16558–16569, 2021. 3, 1
- [72] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 3
- [74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5
- [75] Minh Vo, Ersin Yumer, Kalyan Sunkavalli, Sunil Hadap, Yaser Sheikh, and Srinivasa G Narasimhan. Self-supervised multi-view person association and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2794–2808, 2020. 3
- [76] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13033–13042, 2021. 3
- [77] Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L Yuille, and Wen Gao. Robust estimation of 3d human poses from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2361–2368, 2014. 2
- [78] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, 2024. 1, 3, 4, 5
- [79] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Promptmr: Promptable human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1148–1159, 2025. 1, 3, 4, 6, 7, 2
- [80] Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. Graph-based 3d multi-person pose estimation using multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11148–11157, 2021. 3
- [81] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 2, 4, 8

- [82] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. VITPose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. [2](#), [4](#), [8](#)
- [83] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21222–21232, 2023. [3](#), [7](#)
- [84] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, et al. Whac: World-grounded humans and cameras. In *European Conference on Computer Vision*, pages 20–37. Springer, 2024. [2](#), [3](#)
- [85] Yu Zhan, Fenghai Li, Renliang Weng, and Wongun Choi. Ray3d: ray-based 3d human pose estimation for monocular absolute 3d localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13116–13125, 2022. [3](#)
- [86] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34:13153–13164, 2021. [3](#)
- [87] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13232–13242, 2022. [3](#)
- [88] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. EgoBody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, pages 180–200. Springer, 2022. [2](#)
- [89] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. 2022. [2](#), [4](#)
- [90] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11656–11665, 2021. [3](#)
- [91] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016. [3](#)
- [92] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [5](#)
- [93] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. [3](#), [5](#)