Estimating Egocentric 3D Human Pose in Global Space

Jian Wang[1,2] Lingjie Liu[1,2] Weipeng Xu[3] Kripasindhu Sarkar[1,2] Christian Theobalt[1,2]
[1]MPI Informatics [2]Saarland Informatics Campus [3]Facebook Reality Labs
{jianwang,lliu,ksarkar,theobalt}@mpi-inf.mpg.de xuweipeng@fb.com

Abstract

Egocentric 3D human pose estimation using a single fisheye camera has become popular recently as it allows capturing a wide range of daily activities in unconstrained environments, which is difficult for traditional outside-in motion capture with external cameras. However, existing methods have several limitations. A prominent problem is that the estimated poses lie in the local coordinate system of the fisheye camera, rather than in the world coordinate system, which is restrictive for many applications. Furthermore, these methods suffer from limited accuracy and temporal instability due to ambiguities caused by the monocular setup and the severe occlusion in a strongly distorted egocentric perspective. To tackle these limitations, we present a new method for egocentric global 3D body pose estimation using a single head-mounted fisheye camera. To achieve accurate and temporally stable global poses, a spatio-temporal optimization is performed over a sequence of frames by minimizing heatmap reprojection errors and enforcing local and global body motion priors learned from a mocap dataset. Experimental results show that our approach outperforms state-of-the-art methods both quantitatively and qualitatively.

1. Introduction

Traditional optical motion capture system with external, outside-in facing cameras is restrictive for many pose estimation applications that require the person to be able to roam around in a larger space, beyond a fixed recording volume. Examples are mobile interaction applications, pose estimation in large-scale workplace environments, or many AR/VR applications. To enable this, methods for egocentric 3D human pose estimation using head- or body-mounted cameras were researched. These methods are mobile, flexible, and have the potential to capture a wide range of daily human activities even in large-scale cluttered environments.

Some egocentric capture methods study the estimation of face [9, 8, 21] and hand motions [38, 40, 27, 39], while the estimation of the global full body pose has been less explored. Mo^{2}Cap^2 [45] and xR-egopose [43] use a single head-mounted fisheye camera to capture the 3D skeletal body pose in a marker-less way. Both methods have demonstrated compelling 3D pose estimation results while still suffering from an important limitation: They estimate the local 3D body pose in egocentric camera space, while not being able to obtain the body pose with global position and orientation in the world coordinate system. Henceforth, we will refer to the former as "local pose", in order to distinguish it from the "global pose" defined in the world coordinate system. Local pose capture alone is insufficient for many applications. For example, captured local body poses are not enough to animate the locomotion of a virtual avatar in xR environments, which requires global poses.

A straightforward solution is to simply project the local pose into the world coordinate system with the egocentric camera pose estimated by the SLAM. However, the obtained global poses exhibit significant inaccuracies. First, they show notable temporal jitters as the video frames are processed independently without taking temporal frame coherence. Second, they often show tracking failure due to

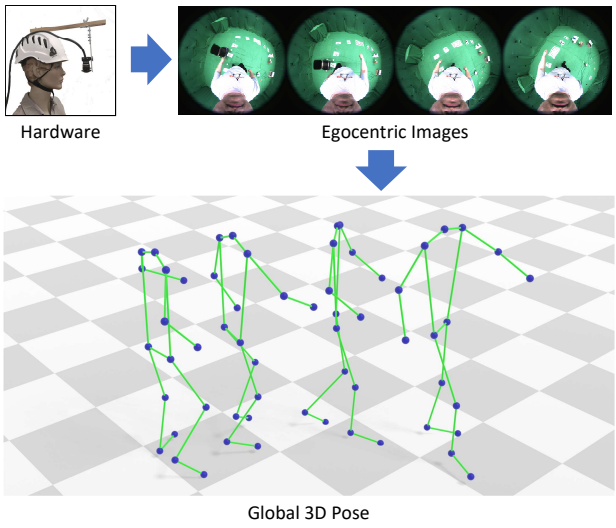

Figure 1. Given challenging egocentric videos, our method produces realistic and accurate 3D global pose sequence.

the self-occlusion in the distorted view of the fisheye camera. Third, the obtained global poses often show unrealistic motions (such as foot sliding and global jitters) due to the inconsistency between the local pose and the estimated camera pose, which are independent of each other.

To tackle these challenges, we propose a novel approach for accurate and temporally stable egocentric global 3D pose estimation with a single head-mounted fisheye camera, as illustrated in Fig. 1. In order to obtain temporally smooth pose sequences, we resort to a spatio-temporal optimization framework where we leverage the 2D and 3D keypoints from CNN detection as well as VAE-based motion priors learned from a large mocap dataset. The VAE-based motion priors have been proven effective to produce realistic and smooth motions in pose estimation methods like VIBE [19] and MEVA[25]. However, the RNN-based VAEs in these works are less efficient and unstable due to the vanishing and exploding gradients during our optimization process. Therefore, we propose a new convolutional VAE-based motion prior, which enables faster optimization speed and higher accuracy. Furthermore, to reduce the error due to strong occlusion, we proposed a novel uncertainty-aware reprojection energy term by summing up the probability values at the pixels on the heatmap occupied by the projection of the 3D estimated joints rather than comparing the projection of 3D estimated joints against the predicted 2D joint position. Finally, in order to make the local body poses consistent with the camera poses estimated by SLAM, we introduce a global pose optimizer with a separate VAE.

We evaluate our method on the dataset provided by Mo^{2}Cap^2 [45] and also a new benchmark we collected with 2 subjects performing various motions. Our method outperforms the state-of-the-art methods both quantitatively and qualitatively. Our ablative analysis confirms the efficacy of our proposed optimization algorithm with learned motion prior and uncertainty-aware reprojection loss for improved local and global accuracy and temporal stability. To summarize, our technical contributions are as follows:

- A novel framework for accurate and temporally stable global 3D human pose estimation from a monocular egocentric video.
- A new optimization algorithm with the assistance of local and global motion prior captured by an efficient convolutional network based VAE.
- An uncertainty-aware reprojection loss to alleviate the influence of self-occlusions in the ego-centric settings.
- State-of-the-art results for local and global pose w.r.t. various baselines.

Our method works for a wide range of motions and daily-life activities in various environments (see the supplemental video in our project page: http://gvv.mpi-inf.mpg.de/projects/globalegomocap/).

2. Related Work

Egocentric 3D full body pose estimation Capturing full-body motion from an egocentric camera perspective has attracted more and more attention in recent years while it is challenging as it is difficult to observe the whole body from close proximity in the egocentric setting. Some works estimate the full-body pose by analyzing the motion of the observed environment. Shiratori *et al*. [37] attach 16 cameras to the subject's limbs and torso to recover the human pose by performing SFM of the environment. Jiang and Grauman [13] reconstruct full-body pose by leveraging learned dynamic and pose coupling over a long time span. Yuan and Kitani [47, 48] use video-conditioned control techniques to estimate and forecast physically-plausible human body motion. Rhodin *et al*. [36] are the first to propose a full-body capture method with a helmet-mounted stereo fisheye camera. Cha *et al*. [5] estimate the 3d body pose from two head-mounted pinhole cameras with a recurrent neural network. To avoid inconvenience of large setup, some researchers use a single wide-view fisheye camera. Xu *et al*. [45] and Tome *et al*. [43] use a compact monocular setting and developed learning-based approaches to estimate ego-pose from a single frame. Hwang *et al*. [11] mount an ultra-wide fisheye camera on the user's chest and estimate body pose, camera rotation and head pose from a single fisheye image. However, these methods neither exploit temporal consistency, nor ensure the reality of predicted motion. Our method, on the contrary, leverages motion prior based optimization approach to make the prediction consistent and accurate.

Leveraging learned prior in 3D pose estimation In order to enhance the accuracy of pose estimation and make predictions more realistic, a lot of recent methods leverage the prior learned from the mocap dataset. Some of them capture the prior in the Gaussian space. For example, Bogo *et al*. [3] and Arnab *et al*. [2] captures the prior to optimize the SMPL body model [24] by fitting a mixture of Gaussians to CMU mocap dataset [1]. Pavlakos *et al*. [34] train a VAE to learn priors of SMPL parameters on AMASS dataset, which contains richer varieties of human motions. Zanfir *et al*. [49] use normalizing flow in order to avoid the compromise between KL divergence and reconstruction loss in VAE. Some other methods incorporate the pose prior by training a generative adversarial network (GAN). Yang *et al*. [46] develop a adversarial learning framework with multi-source discriminator. Kanazawa *et al*. [14, 15] and Zhang *et al*. [51] train discriminators for each joint rotation parameter to tell if these parameters are realistic. Kocabas *et al*. [19] propose a temporal network architecture with an RNN-based discriminator for the adversarial training on the sequence of SMPL parameters. Different from previous methods, our method captures the global motion

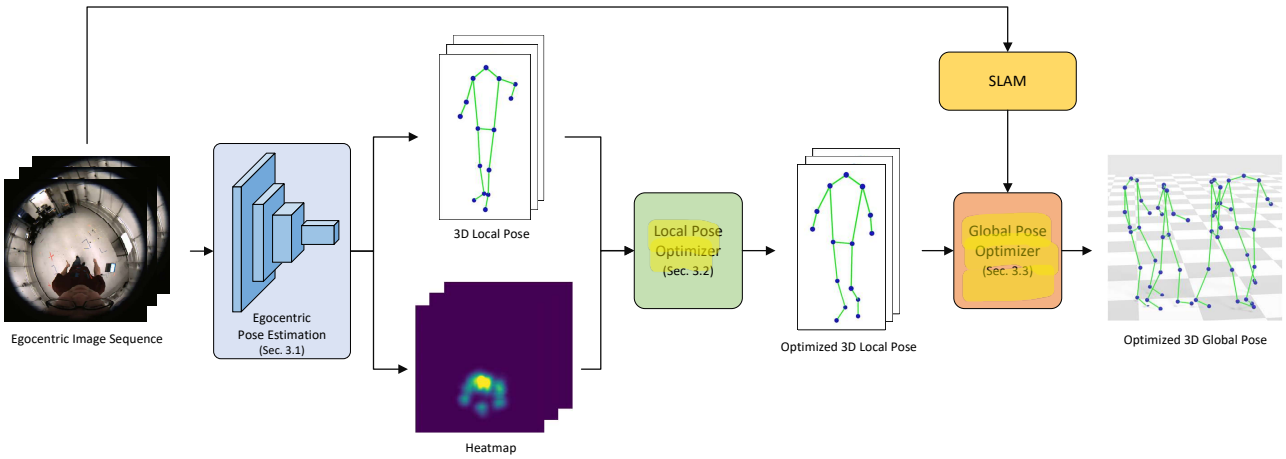

Figure 2. Overview of our method. Our method takes an egocentric video as input and processes it in segments. For each segment consisting of a fixed number of consecutive frames, we first apply an egocentric pose estimation method to obtain initial 3D local poses and 2D heatmaps which are then fed into the local pose optimization framework to get optimized local poses. Next, combined with the camera poses estimated from ORB-SLAM2, the optimized 3D local poses are transformed from the local egocentric camera space to the world coordinate space and then optimized via the global pose optimization to produce the final global poses.

prior learned with a light-weight sequential VAE, which enables direct optimization in the global coordinate system.

Monocular 3D pose estimation in video Monocular 3D pose estimation has been the focus of research for a long time. Some methods predict 2D joints and perform 2D-to-3D lifting separately [6, 12, 29], while some other methods regress the 3D pose directly [22, 30, 41, 42, 20]. These methods process single image and therefore exhibit notable temporal jitter in a video sequence. To solve this, many recent methods exploit temporal information from the video. Zhou *et al*. [52] introduce EM method to estimate 3D pose from 2D predictions over the entire sequence. Mehta *et al*. [32] and Du *et al*. [7] apply temporal filtering across 2D and 3D poses. Lin *et al*. [23], Hossain *et al*. [10], Kocabas *et al*. [19] and Katircioglu *et al*. [16] use recurrent networks to predict 3D pose sequences by leveraging previously predicted 2D and 3D poses. Pavllo *et al*. [35] generates 3D poses with temporal-convolution, while Cai *et al*. [4] and Wang *et al*. [44] leverage graph convolutional network to capture the temporal information. Luo *et al*. [26] firstly get coarse motion with a GRU based human motion VAE and then refine the motion with a residual estimation network. Different from all previous works, our method capture the motion prior with a 1D convolution based sequential VAE, and we use the VAE in our optimization framework.

3. Method

Our goal is to estimate the global body poses from a video sequence captured by a head-mounted fisheye camera. We provide an overview of our pipeline in Fig. 2. The video frames are split into segments with B frames each ($B = 10$ in our experiments). Our pipeline takes one segment consisting of B consecutive frames, $\mathcal{I}_{seq} = \{\mathcal{I}_1, \dots, \mathcal{I}_B\}$, as inputs and outputs the global poses of all the individual frames, $\mathcal{P}^g_{seq} = \{\mathcal{P}^g_1, \dots, \mathcal{P}^g_B\}$. For each segment, we first calculate the 3D local pose and 2D heatmap of each frame using an egocentric local body pose estimation method (Sec. 3.1). Next, we learn the local motion prior from local motion sequences of the AMASS dataset [28] with a sequential VAE [18] (Sec. 3.2.1), and perform a spatio-temporal optimization with the local motion prior by minimizing the heatmap reprojection term and several regularization terms (Sec. 3.2.2). Given the optimized local poses, we transform them from local fisheye camera space to the world coordinate system with camera poses estimated by a SLAM method to get initial global poses (Sec. 3.3.1). To improve global poses, we learn the global pose prior by training a second sequential VAE on the global motion sequences of the AMASS dataset, and impose the global pose prior in a spatio-temporal global pose optimization (Sec. 3.3.2). Please refer to the supplementary materials for our implementation details.

3.1. Local Pose Estimation

Given a segment containing B consecutive frames $\mathcal{I}_{seq}$, we estimate local poses represented by 15 joint locations $\widetilde{\mathcal{P}}_{seq} = \{\widetilde{\mathcal{P}}_1, \dots, \widetilde{\mathcal{P}}_B\}$, $\widetilde{\mathcal{P}}_i \in \mathbb{R}^{15\times 3}$, and 2D heatmaps $\mathcal{H}_{seq} = \{\mathcal{H}_1, \dots, \mathcal{H}_B\}$ using an egocentric local pose estimation method. Note that our approach can work with any egocentric local pose estimation methods. In our experiments, we evaluate our approach on the results of two state-of-the-art methods: Mo^{2}Cap^2 [45] and xR-egopose [43].

3.2. Local Pose Optimization

Although Mo^{2}Cap^2 and xR-egopose can produce compelling results, both approaches suffer from limited accuracy and temporal instability, which is mainly due to depth ambiguities caused by the monocular setup and severe occlusions in a strongly distorted egocentric perspective. To improve local poses, we design an efficient spatio-temporal optimization framework which first learns the local pose prior as a latent space with a sequential VAE [18] (Sec. 3.2.1) and then searches for a latent vector in the learned latent space by minimizing a reprojection term and some regularization terms (Sec. 3.2.2).

3.2.1 Learning Motion Prior

To construct a latent space encoding local motion prior, we train a sequential VAE [18] on local motion sequences of the AMASS dataset [28] which are split into segments for training. We denote a segment consisting of B consecutive poses as $\mathcal{Q}_{seq} = \{\mathcal{Q}_1, \dots, \mathcal{Q}_B\} (\mathcal{Q}_i \in \mathbb{R}^{15\times 3})$. The sequential VAE consists of an encoder f_{enc} and a decoder f_{dec}. The encoder is used to map an input sequence of human local poses $\mathcal{Q}_{seq}$ to a latent vector z, and the decoder is used to reconstruct a pose sequence, $\widehat{\mathcal{Q}}_{seq} = \{\widehat{\mathcal{Q}}_1, \dots, \widehat{\mathcal{Q}}_B\} (\widehat{\mathcal{Q}}_i \in \mathbb{R}^{15\times 3})$, from the latent vector. Following [18], the training loss of VAE is formulated as:

$$\begin{aligned} \mathcal{L}_{total} = c_1 \left\| \widehat{\mathcal{Q}}_{seq} - \mathcal{Q}_{seq} \right\|_2^2 \\ + c_2 KL[q(z|\mathcal{Q}_{seq}) \| \mathcal{N}(0, I)] \end{aligned} \tag{1}$$

where $z = f_{enc}(\mathcal{Q}_{seq})$, $\widehat{\mathcal{Q}}_{seq} = f_{dec}(z)$, $q(z|\mathcal{Q}_{seq})$ refers to the projected distribution of $\mathcal{Q}_{seq}$ in the latent space, $\mathcal{N}(0, I)$ refers to the standard normal distribution, and $KL(.)$ refers to the Kullback–Leibler divergence.

Different from previous pose estimation methods [19, 25] which leverage RNN-based VAEs to capture the motion prior, both the encoder f_{enc} and the decoder f_{dec} of our sequential VAE are designed as 5-layer 1D convolutional networks. Comparing with RNN-based VAEs, the convolutional networks in our sequential VAE is more efficient in the optimization iterations since it can be parallelized over time sequence. Moreover, the RNNs suffer from vanishing and exploding gradients more easily, which makes optimization process less stable. We have compared the sequential VAE in our method with RNN-based VAEs in VIBE [19] and MEVA [25] in Sec. 4.4. More details of sequential VAE is shown in the supplementary materials.

3.2.2 Optimizing Local Poses with Local Motion Prior

With the learned latent space of local motion, the task of optimizing local poses with the local motion prior can be formulated as the problem of finding a latent vector z in the learned latent space such that the reconstructed local pose sequence $\mathcal{P}_{seq} = f_{dec}(z)$ minimizes the following objective function:

$$\begin{aligned} E(\mathcal{P}_{seq}) = \lambda_R E_R(\mathcal{P}_{seq}) + \lambda_J E_J(\mathcal{P}_{seq}, \widetilde{\mathcal{P}}_{seq}) \\ + \lambda_T E_T(\mathcal{P}_{seq}) + \lambda_B E_B(\mathcal{P}_{seq}) \end{aligned} \tag{2}$$

where $E_R(.), E_J(.), E_T(.), E_B(.)$ are the reprojection term, pose regularization term, motion smoothness regularization term and bone length regularization term, respectively, which we will describe in detail later. In our experiment, we set the weights $\lambda_R = 0.01$, $\lambda_J = 0.01$, $\lambda_T = 1$ and $\lambda_B = 0.01$, respectively.

Heatmap-based Reprojection: Previous works [2, 3, 34, 50] calculate the reprojection term by summing up the Euclidean distance values between the projection of estimated 3D joints and detected 2D joints. However, this calculation is sensitive to 2D joint detection errors due to the strong self-occlusions caused by the egocentric perspective. To tackle this issue, we define a heatmap-based reprojection error by leveraging the uncertainty captured in the predicted 2D heatmaps, where the value at each pixel describes the probability of this pixel being a 2D joint. This new reprojection term is calculated by maximizing the summed heatmap values at the reprojected 2D joint positions:

$$E_R(\mathcal{P}_{seq}) = -\sum_{i=1}^{B} \|\text{HM}_i(\Pi(\mathcal{P}_i))\|_2^2 \tag{3}$$

where $\text{HM}_i(.)$ returns the value at a pixel on $\mathcal{H}_i$, the heatmap of i-th frame. $\Pi(.)$ refers to the projection of a 3D point. Specifically, the projection of a 3D point $[x, y, z]^T$ can be written as:

$$[u, v]^T = \frac{[x, y]^T}{\sqrt{x^2 + y^2}} \times f(\rho) \tag{4}$$

where $\rho = \arctan(z/\sqrt{x^2 + y^2})$ and $f(\rho) = \alpha_0 + \alpha_1\rho + \alpha_2\rho^2 + \alpha_3\rho^3 + \dots$ is a polynomial obtained from camera calibration.

Pose Regularization: To constrain the optimized pose $\mathcal{P}_i$ to stay close to the initial pose $\widetilde{\mathcal{P}}_i$, we define the pose regularizer as:

$$E_J(\mathcal{P}_{seq}, \widetilde{\mathcal{P}}_{seq}) = \sum_{i=1}^{B} \left\| \mathcal{P}_i - \widetilde{\mathcal{P}}_i \right\|_2^2 \tag{5}$$

Motion Smoothness Regularization: Same as [31], the temporal smoothness regularizer (Eq. 6) is used to improve

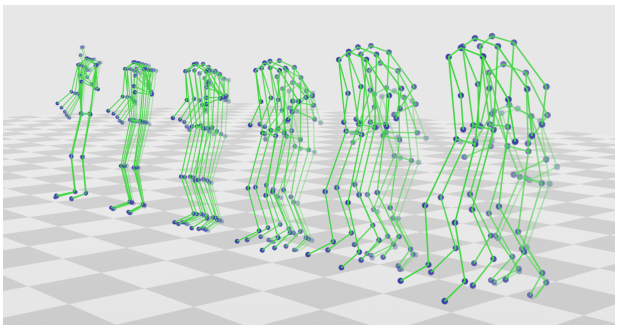

Figure 3. Interpolation in the latent space. The leftmost and rightmost pose sequences (waving hands and jumping) are reconstructed from two randomly sampled latent vectors, and intermediate pose sequences are reconstructed from linear interpolation between the left and right latent vectors.

the temporal stability of the estimated poses, which is calculated based on the acceleration of each joint over the whole sequence:

$$E_T(\mathcal{P}_{seq}) = \sum_{i=2}^{B} \|\nabla\mathcal{P}_i - \nabla\mathcal{P}_{i-1}\|_2^2 \tag{6}$$

where $\nabla\mathcal{P}_i = \mathcal{P}_i - \mathcal{P}_{i-1}$.

Bone Length Regularization: To explicitly enforce the constraint that each bone length stay fixed, we define the bone length regularizer as the difference between the bone length and the average bone length over the pose sequence.

$$E_B(\mathcal{P}_{seq}) = \sum_{i=1}^{B} \left\| L_{\mathcal{P}_i} - \frac{1}{B}\sum_{j=1}^{B} L_{\mathcal{P}_j} \right\|_2^2 \tag{7}$$

where the $L_{\mathcal{P}_i}$ is a vector composed of the length of each bone of 3D pose $\mathcal{P}_i$.

3.3. Global Pose Estimation

Based on the pose optimized by the local pose optimizer, we seek to get the 3D pose in the global coordinate system. We firstly use the monocular SLAM to get the camera pose sequence and project the local pose sequence to the global space (Sec. 3.3.1), then we optimize the initial global pose sequence with our global pose optimizer (Sec. 3.3.2).

3.3.1 Initialization

To obtain the initial global body poses, we first estimate the camera poses using ORB-SLAM2 [33]. In order to avoid the effects caused by the moving person in the egocentric view, we employ a square-shaped mask that roughly covers a large portion of the body to remove most of the feature points detected on the main body parts. We use a fixed mask rather than estimating a silhouette mask for each image for the sake of effectiveness and robustness.

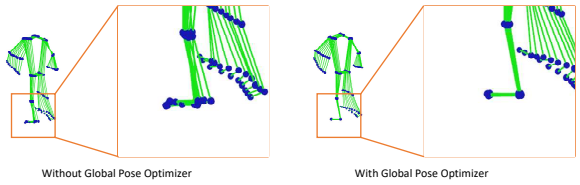

Figure 4. The global pose with/without global pose optimizer. The left foot is zoomed in for better comparison.

With the estimated camera pose (R_i, t_i) $(i = 1, \cdots, B)$, the local body pose P_i can be transformed into the world coordinate space to obtain its initial global body pose $\widetilde{P}_i^g$:

$$\widetilde{\mathcal{P}}_i^g = R_i \cdot \mathcal{P}_i + t_i, \widetilde{\mathcal{P}}_i^g \in \widetilde{\mathcal{P}}_{seq}^g \tag{8}$$

where $\widetilde{\mathcal{P}}_{seq}^g$ is the corresponding inital global pose segment of $\mathcal{P}_{seq}$.

3.3.2 Global Pose Optimizer

Simply combining local poses with camera poses would not achieve very high-quality global poses because the optimized local body poses are not constrained to be consistent with the corresponding camera poses. For example, the initial global pose in the left part of Fig. 4 suffers from the footskate artifact, which means the foot moves when it should remain in a fixed position on the ground. In order to alleviate such inconsistency errors, we perform another spatio-temporal optimization on the initial global pose. We first train a sequential VAE on global pose sequences from the AMASS dataset in the same way presented in Sec. 3.2.1. To measure the smoothness of our learned latent space, we conducted an experiment of interpolating two different body motions. The results shown in Fig. 3 demonstrate that the learned latent space is smooth (also see this result in the supplemental video), which is important for the subsequent optimization process. With the learned latent space of global motion, we seek for a latent vector z^g such that the global pose sequence $\mathcal{P}_{seq}^g = f_{dec}^g(z^g)$ minimizes the following objective function:

$$\begin{aligned} E(\mathcal{P}_{seq}^g) = \lambda_J E_J(\mathcal{P}_{seq}^g, \widetilde{\mathcal{P}}_{seq}^g) + \lambda_T E_T(\mathcal{P}_{seq}^g) \\ + \lambda_B E_B(\mathcal{P}_{seq}^g) \end{aligned} \tag{9}$$

where $E_J(.), E_T(.), E_B(.)$ are the same as those in 3.2.2, and λ_J, λ_T and λ_B are set as 0.01, 1 and 0.01, respectively. The example of optimized result is illustrated in the right part of Fig. 4, where the footskate artifact is alleviated due to our global optimizer.

4. Experiments

4.1. Datasets

Following [45] and [43], we train our local egocentric pose estimators on the synthetic dataset from Mo^{2}Cap^2. We use the AMASS dataset [28] to train our sequential VAEs. To make the distribution of joint position in the training data consistent with that in the real-world data, we set a virtual fisheye camera attached to the forehead of the human mesh at a distance similar to our capture settings, thus we are able to calculate the local and the global pose in the fisheye coordinate system for training both the VAEs.

We evaluate our method on both the real-world dataset from Mo^{2}Cap^2 [45] and a new egocentric dataset. Our new real-world dataset was captured using a head-mounted fisheye camera with the similar camera position as Mo^{2}Cap^2 [45] while the ground truth 3D poses were acquired using a multi-view motion capture system. This dataset contains around 12k frames of 2 actors wearing different clothes and performing 13 types of actions. This dataset will be made publicly available and further details of it are shown in the supplementary materials.

4.2. Evaluation Metrics

We evaluate our method with three different metrics, namely PA-MPJPE, the bone length aligned MPJPE (BA-MPJPE) and the global MPJPE. They all calculate the Mean Per Joint Position Error (MPJPE) but use different ways of alignment to the ground truth. For **PA-MPJPE**, we rigidly align the estimated pose of each frame to the ground truth pose P_{seq} using $\hat{P}_{seq}$ with Procrustes analysis [17]. For **BA-MPJPE**, we first resize the bone length of each frame in sequences $\hat{P}_{seq}$ and P_{seq} to the bone length of a standard skeleton. Then, we calculate the PA-MPJPE between the two resulting sequences. For **Global MPJPE**, we globally align all the poses of each batch (100 frames) to the ground truth using Procrustes analysis.

Each of the three metrics has its own focus. The PA-MPJPE measures the accuracy of a single pose while BA-MPJPE eliminates the effects of body scale. The global MPJPE calculates the accuracy of global joint positions, considering the global translation and rotation.

4.3. Comparison with State-of-the-art Results

Table 1 compares our approach with previous state-of-the-art single-frame-based methods on our dataset and the indoor sequence of Mo^{2}Cap^2 dataset. Since the code or the predictions of xR-egopose are not publicly available, we use our implementation instead. In order to obtain the global pose for Mo^{2}Cap^2 and xR-egopose, we rigidly transform the local predictions to the world coordinate system with the camera pose estimated by SLAM. This global pose is regarded as our main baseline and denoted as Mo^{2}Cap^2 (or xR-egopose) + SLAM. Since the camera poses from ORB-SLAM2 are ambiguous to the scene scale, we further estimate the scale by calibrating the camera position with a checkerboard in the first few frames of the sequence. Note that since the Mo^{2}Cap^2 dataset does not provide frames with a checkerboard, we applied the Procrustes analysis to align the trajectory estimated by SLAM with the ground truth trajectory to compute the scale. For a fair comparison, we also smoothed the global pose of Mo^{2}Cap^2 and xR-egopose with a Gaussian filter and denote the results as Mo^{2}Cap^2 (or xR-egopose) + SLAM + smooth.

Method	Global MPJPE	PA-MPJPE	BA-MPJPE
Mo^{2}Cap^2 test dataset			
Mo^{2}Cap^2+SLAM	117.4	80.48	61.40
Mo^{2}Cap^2+SLAM+Smooth	113.0	76.92	58.25
Mo^{2}Cap^2+Ours	**106.9**	**66.95**	**50.20**
xR-egopose+SLAM	114.0	71.33	55.43
xR-egopose+SLAM+Smooth	112.2	70.27	54.03
xR-egopose+Ours	**108.2**	**64.15**	**48.34**
Our test dataset			
Mo^{2}Cap^2+SLAM	141.8	102.3	74.46
Mo^{2}Cap^2+SLAM+Smooth	135.5	96.37	70.84
Mo^{2}Cap^2 + Ours	**117.4**	**79.44**	**61.38**
xR-egopose+SLAM	163.4	112.0	87.20
xR-egopose+SLAM+Smooth	158.1	109.6	84.70
xR-egopose+Ours	**127.7**	**82.83**	**64.20**

Table 1. The experimental results on Mo^{2}Cap^2 test dataset [45] and our test dataset. Mo^{2}Cap^2 (or xR-egopose) + Ours is the result of our method based on the predictions of Mo^{2}Cap^2 (or xR-egopose). Our method outperforms previous state-of-the-art Mo^{2}Cap^2 [45] and xR-egopose [43] in all of the three metrics.

From all the aforementioned comparisons, we observe significant improvements, which proves that our method is able to improve the accuracy of pose estimation results from egocentric videos. Please also refer to the supplementary materials for the BA-MPJPE on each type of motion. For the qualitative evaluation, we show the comparison between Mo^{2}Cap^2 and our method (based on Mo^{2}Cap^2) in Fig. 5. Please also see our supplementary video for more results. Our method also features the ability to estimate the global body pose, which is shown in Fig. 6 and our supplementary video. In Fig. 6 we demonstrate the accuracy of our global pose estimation by projecting the predicted global pose to an external camera.

4.4. Ablation Study

We further conduct experiments to evaluate the effects of individual components of our approach. We use Mo^{2}Cap^2 as our local pose estimator for all our ablation studies to make the results comparable.

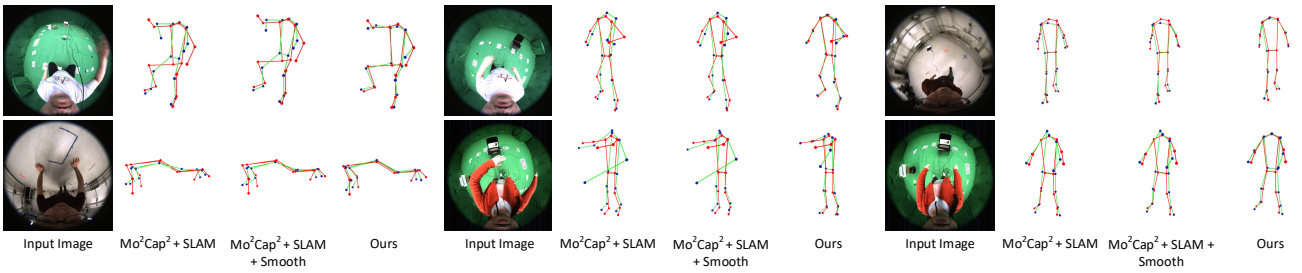

Figure 5. Qualitative comparison on the accuracy of a single pose. From left to right: input image, Mo^2Cap^2 result projected with SLAM (green), smoothed Mo^2Cap^2 result projected with SLAM (green) and our result (green) overlaid on ground truth (red). Note that in order to better show the result, we rigidly align the estimated pose to the ground truth.

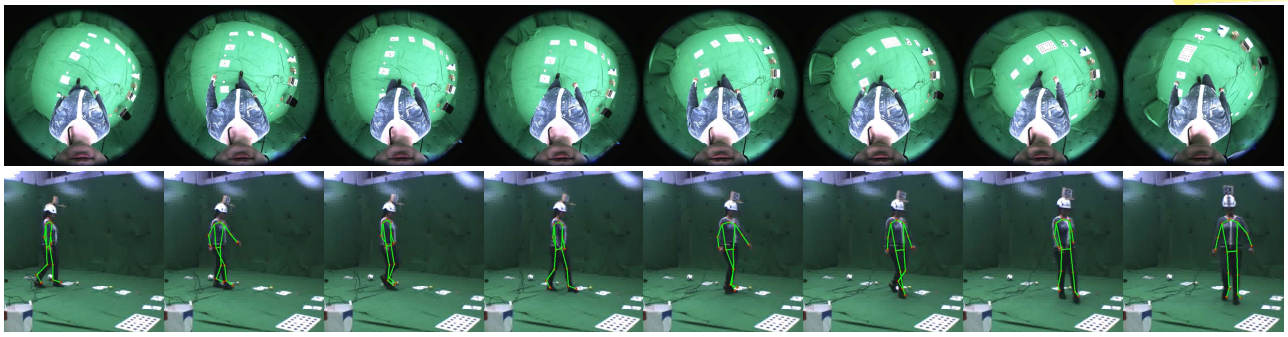

Figure 6. Global pose estimation results from a third-view camera. Top row: the input egocentric images, bottom row: the estimated 3D pose projected on an external camera.

Method	Global MPJPE	PA-MPJPE	BA-MPJPE
Mo^2Cap^2 + SLAM	141.8	102.3	74.46
w/o local optim.	128.8	89.76	68.72
w/o global optim.	123.9	84.18	63.96
w/o motion prior	127.3	90.71	68.78
w. GMM	124.2	90.00	67.69
w. single frame VAE	124.7	89.88	67.86
w. VAE in VIBE	123.1	83.37	65.49
w. VAE in MEVA	120.7	81.57	63.32
w. MLP based VAE	120.5	82.55	63.49
conventional reproj.	125.9	86.89	66.92
Mo^2Cap^2 + Ours	**117.4**	**79.44**	**61.38**

Table 2. The quantitative results of ablation study.

Local/ global pose optimizer. In this experiment, in order to investigate the efficacy of our local and global optimizer, we evaluate our method after removing the local pose optimizer or the global pose optimizer from our whole pipeline. The results are shown in the 2nd and 3rd row of Table 2, which shows that both of the modules are important to our approach. The heatmap reprojection error in the local pose optimizer ensures that the optimized 3D pose conforms to the constraint of 2D predictions. The VAE prior in the global pose optimizer keeps the movement of body limbs in accordance with the global camera pose, thus improves both on the global MPJPE and the local MPJPEs.

Motion priors. In order to validate the importance of motion priors, we test the performance of our optimization framework without our motion priors by directly optimizing 3D pose P_{seq} with $E(P_{seq})$ rather than optimizing the VAE's latent vector z. We evaluate the method without motion prior on our dataset and show one of our results in Fig. 7. In this figure, the human leg in the input image is severely occluded. The ambiguity of the image significantly reduced the accuracy of our single-frame pose estimation network. Without the motion prior, our optimization framework cannot resolve the ambiguity and the error is still large, while in our method, the motion prior is able to correct the estimated pose. The qualitative evaluation in the 4th row of Table 2 also confirms our claim. With the motion prior, our spatio-temporal optimization framework is able to make pose predictions smoother and less ambiguous.

We also compared our prior with the gaussian mixture model (GMM) prior used in [3, 2, 20] and the single-frame VAE prior used in [34]. When comparing with GMM prior, we firstly train the GMM model with 8 Gaussians on the local pose sequence (local GMM) and the global pose sequence (global GMM) from the AMASS dataset. Then we

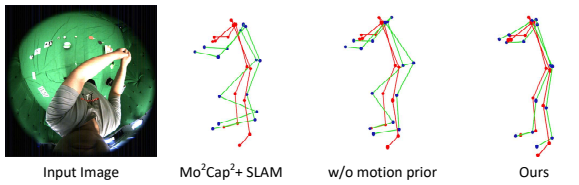

Figure 7. Comparison between our method with and without motion prior. From left to right: input image, Mo^{2}Cap^2 + SLAM (green), the result without motion prior (green) and the one with motion prior (our result) (green) overlaid on the ground truth (red).

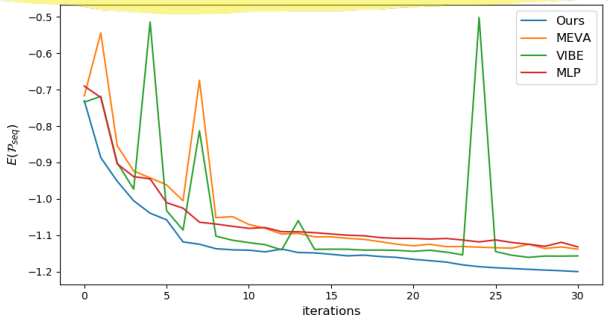

Figure 8. $E(\mathcal{P}_{seq})$-iteration curve of different VAEs. Our method gives the lowest error while keeping stable during optimization.

substitute the local and global VAE in our method with the local and global GMM and evaluate three MPJPEs, which is shown in the 5th row of Table 2. GMM prior performs worse since the VAE uses the neural network as a feature extractor, making it easier to capture priors. When comparing with single-frame based VAE prior, we train a VAE network taking a single input pose on the AMASS dataset and substitute the VAE in the local optimizer with the single-frame VAE. The evaluation result is shown in the 6th row of Table 2. The single-frame VAE cannot capture the prior over time, making it less effective than our sequential VAE.

CNN based sequential VAE. We use the CNN-based sequential VAE rather than RNN-based VAE for better efficiency and optimization stability. To evaluate our advantage, we substitute our CNN-based sequential VAE in both the local and global optimizer with the VAEs in VIBE [19] or MEVA [25] (see supplementary materials for implementation details), and report the results in the 7th to 9th rows of Table 2. The result proves that our CNN-based VAE outperforms others in terms of optimization accuracy, which can be attributed to a more stable optimization process. To demonstrate this, we show the the $E(\mathcal{P}_{seq})$-iteration curve of local pose optimization process (Sec. 3.2.2) in Fig. 8, where RNN-based VAEs are less stable due to the gradient explosion issue. To show the efficiency of CNN-based VAE, we evaluated the time needed for the optimization. Our method takes 195.7ms per 10-frame segment while

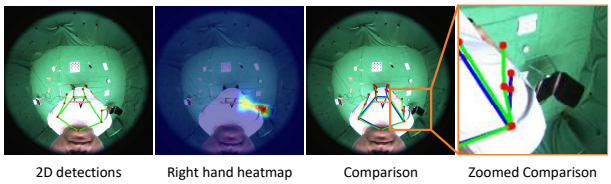

Figure 9. Comparison between heatmap reprojection error and conventional reprojection error. In the 3rd and 4th image from left, we show the result of heatmap reprojection error in green skeleton and result of conventional reprojection error in blue skeleton.

RNN-based VAE in VIBE and MEVA takes 552.1ms and 1139.4ms per segment respectively. We also compared our CNN-based VAE with multilayer perceptron (MLP) based VAE. According to Fig. 8 and the 10th row of Table 2, the MLP-based VAE performs worse since it is not designed to capture the temporal context of the pose sequence.

Heatmap reprojection error. In this work we use the heatmap reprojection error while a lot of previous works get the reprojection error by calculating the distance between estimated 2D joints and corresponding projected 3D joints [2, 3, 34, 50]. To evaluate the improvement of heatmap reprojection error over the previous approach, we substitute the heatmap reprojection error in our pipeline with the conventional reprojection error in [3] and compare this with our method. In the qualitative evaluation shown in Fig. 9, the 2D pose estimation gives wrong results for the right-hand position while the ground truth hand position is still covered by the heatmap. Our heatmap reprojection error can leverage such uncertainty in the heatmap and gives better results than the conventional reprojection error. We also show the quantitative result in the 10th row of Table 2. These results validate the advantage of our heatmap reprojection error.

5. Conclusions

In this paper, we propose a method for estimating global poses with a single head-mounted fisheye camera. This is achieved by employing novel strategies in our spatio-temporal optimization framework: (1) a sequential VAE to effectively capture the body motion prior. (2) a global motion prior to ensure consistency between the local body motion and the camera poses. (3) a heatmap-based reprojection error term to leverage the uncertainty in predicted heatmaps. Extensive experiments show that our method outperforms state-of-the-art methods. We further evaluate the effects of individual components of our approach.

In future work, we will study the solutions to this problem such as the integration of depth sensors. Other future research directions include using the optimized 3D pose in real world to finetune the local pose estimation network and applying our method to the multi-person scenario.

Acknowledgments

Jian Wang, Kripasindhu Sarkar and Christian Theobalt have been supported by the ERC Consolidator Grant 4DReply (770784) and Lingjie Liu has been supported by Lise Meitner Postdoctoral Fellowship.

References

[1] CMU mocap dataset. http://mocap.cs.cmu.edu/, 2008. 2

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 2, 4, 7, 8

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. 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*, volume 9909, pages 561–578, 2016. 2, 4, 7, 8

[4] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat-Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *IEEE International Conference on Computer Vision*, pages 2272–2281, 2019. 3

[5] Y. Cha, T. Price, Z. Wei, X. Lu, N. Rewkowski, R. Chabra, Z. Qin, H. Kim, Z. Su, Y. Liu, A. Ilie, A. State, Z. Xu, J. Frahm, and H. Fuchs. Towards fully mobile 3d face, body, and environment capture using only head-worn cameras. *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2993–3004, 2018. 2

[6] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5759–5767, 2017. 3

[7] Yu Du, Yongkang Wong, Yonghao Liu, Feilin Han, Yilin Gui, Zhen Wang, Mohan S. Kankanhalli, and Weidong Geng. Marker-less 3d human motion capture with monocular image sequence and height-maps. In *European Conference on Computer Vision*, volume 9908, pages 20–36, 2016. 3

[8] Mohamed Elgharib, Mallikarjun BR, Ayush Tewari, Hyeongwoo Kim, Wentao Liu, Hans-Peter Seidel, and Christian Theobalt. Egoface: Egocentric face performance capture and videorealistic reenactment, 2019. 1

[9] Mohamed Elgharib, Mohit Mendiratta, Justus Thies, Matthias Nießner, Hans-Peter Seidel, Ayush Tewari, Vladislav Golyanik, and Christian Theobalt. Egocentric videoconferencing. *ACM Transactions on Graphics*, 39(6), Dec 2020. 1

[10] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision*, volume 11214, pages 69–86, 2018. 3

[11] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. Monoeye: Multimodal human motion capture system using a single ultra-wide fisheye camera. In *ACM Symposium on User Interface Software & Technology*, page 98–111, 2020. 2

[12] Ehsan Jahangiri and Alan L. Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *IEEE International Conference on Computer Vision Workshops*, pages 805–814, 2017. 3

[13] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3509, 2017. 2

[14] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2

[15] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 2

[16] Isinsu Katircioglu, Bugra Tekin, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Learning latent representations of 3d human pose with deep neural networks. *Int. J. Comput. Vis.*, 126(12):1326–1341, 2018. 3

[17] David G Kendall. A survey of the statistical theory of shape. *Statistical Science*, pages 87–99, 1989. 6

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. 3, 4

[19] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5252–5262, 2020. 2, 3, 4, 8

[20] 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 *IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 3, 7

[21] Hao Li, Laura C. Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. Facial performance sensing head-mounted display. *ACM Trans. Graph.*, 34(4):47:1–47:9, 2015. 1

[22] Sijin Li and Antoni B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, volume 9004, pages 332–347, 2014. 3

[23] Mude Lin, Liang Lin, Xiaodan Liang, Keze Wang, and Hui Cheng. Recurrent 3d pose sequence machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5543–5552, 2017. 3

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2

[25] 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 (ACCV)*, November 2020. 2, 4, 8

[26] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. *arXiv preprint arXiv:2008.03789*, 2020. 3

[27] Minghuang Ma, Haoqi Fan, and Kris M. Kitani. Going deeper into first-person activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016. 1

[28] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *IEEE International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 3, 4, 6

[29] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision*, pages 2659–2668, 2017. 3

[30] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved CNN supervision. In *International Conference on 3D Vision*, pages 506–516, 2017. 3

[31] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: real-time multi-person 3d motion capture with a single RGB camera. *ACM Trans. Graph.*, 39(4):82, 2020. 4

[32] 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 Trans. Graph.*, 36(4):44:1–44:14, 2017. 3

[33] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 5

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2, 4, 7, 8

[35] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 3

[36] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Trans. Graph.*, 35(6):162:1–162:11, 2016. 2

[37] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K. Hodgins. Motion capture from body-mounted cameras. *ACM Trans. Graph.*, 30(4):31, 2011. 2

[38] S. Singh, C. Arora, and C. V. Jawahar. First person action recognition using deep learned descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2620–2628, 2016. 1

[39] Suriya Singh, Chetan Arora, and C. V. Jawahar. Trajectory aligned features for first person action recognition. *Pattern Recognit.*, 62:45–55, 2017. 1

[40] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 1

[41] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. In *British Machine Vision Conference*, 2016. 3

[42] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Fusing 2d uncertainty and 3d cues for monocular body pose estimation. *arXiv preprint arXiv:1611.05708*, 2(3), 2016. 3

[43] Denis Tomè, Patrick Peluse, Lourdes Agapito, and Hernán Badino. xr-egopose: Egocentric 3d human pose from an HMD camera. In *IEEE International Conference on Computer Vision*, pages 7727–7737, 2019. 1, 2, 3, 6

[44] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. *arXiv preprint arXiv:2004.13985*, 2020. 3

[45] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo^2cap^2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE Trans. Vis. Comput. Graph.*, 25(5):2093–2101, 2019. 1, 2, 3, 6

[46] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy S. J. Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5255–5264, 2018. 2

[47] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *European Conference on Computer Vision*, pages 735–750, 2018. 2

[48] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time PD control. In *IEEE International Conference on Computer Vision*, pages 10081–10091, 2019. 2

[49] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. *arXiv preprint arXiv:2003.10350*, 2020. 2

[50] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision*, volume 12351, pages 465–481, 2020. 4, 8

[51] Jason Y. Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *IEEE International Conference on Computer Vision*, pages 7113–7122, 2019. 2

[52] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4966–4975, 2016. 3