WorldPose: A World Cup Dataset for Global 3D Human Pose Estimation

Tianjian Jiang[*1], Johsan Billingham[*2], Sebastian Mükesch[1], Juan Zarate[1], Nicolas Evans[2], Martin R. Oswald[3], Marc Polleyfeys[1,4], Otmar Hilliges[1], Manuel Kaufmann[1], and Jie Song[†1]

[1] ETH Zürich
[2] FIFA
[3] University of Amsterdam
[4] Microsoft

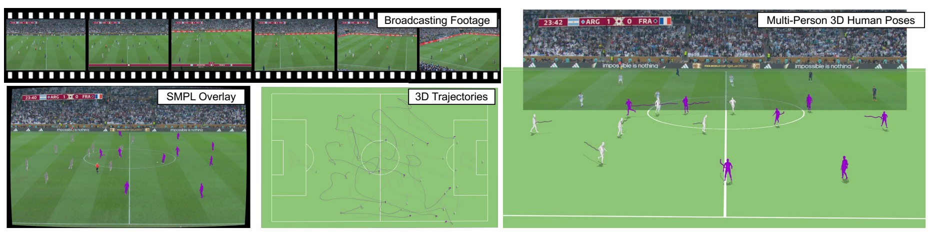

Fig. 1: We leverage multi-view cameras to curate WorldPose, a comprehensive dataset designed for multi-person 3D human pose estimation with global trajectories.

Abstract. We present WorldPose, a novel dataset for advancing research in multi-person global pose estimation in the wild, featuring footage from the 2022 FIFA World Cup. While previous datasets have primarily focused on local poses, often limited to a single person or in constrained, indoor settings, the infrastructure deployed for this sporting event allows access to multiple fixed and moving cameras in different stadiums. We exploit the static multi-view setup of HD cameras to recover the 3D player poses and motions with unprecedented accuracy given capture areas of more than 1.75 acres (7k m^2). We then leverage the captured players' motions and field markings to calibrate a moving broadcasting camera. The resulting dataset comprises 88 sequences with more than 2.5 million 3D poses and a total traveling distance of over 120 km. Subsequently, we conduct an in-depth analysis of the SOTA methods for global pose estimation. Our experiments demonstrate that WorldPose challenges existing multi-person techniques, supporting the potential for new research in this area and others, such as sports analysis. All pose annotations (in SMPL format), broadcasting camera parameters and footage will be released for academic research purposes.

Keywords: human pose estimation · multi-person pose estimation · global trajectory estimation

[*] These authors contributed equally to this work
[†] Now at HKUST(GZ)&HKUST

1 Introduction

The analysis of social interactions, crowd behavior, and team dynamics of human groups offers valuable insights for sociological research in various domains. For instance, in sports science, precise 3D pose analysis for team activities could be a game-changing tool to improve training strategies, prevent injuries, and optimize overall team performance. In recent years, significant advancements have been achieved in the development of Computer Vision and Deep Learning techniques for human pose estimation [3, 17, 26, 33, 36, 37, 42, 54]. However, existing pose estimation datasets used for the evaluation of these methods are predominantly designed for single individuals [18, 23, 27, 28, 40, 50, 69]. Those datasets that do feature multi-person scenarios are often constrained to lab-based settings [12,15, 25,61,66], resulting in a limited number of individuals or restricted movement due to spatial constraints [22,29,39,41]. However, many real-world scenarios involve large groups of people engaging in coordinated, dynamic activities that can occur in open, expansive outdoor areas and often involve moving cameras. Hence, these datasets are inadequate for capturing the complexity of these scenarios and especially insufficient for understanding the relative positions of multiple individuals over long trajectories.

Capturing a dataset that addresses these problems is a major challenge. First, the vast capture area renders marker-based methods relying on IR reflection impractical. Second, solutions based on body-worn sensors often encounter significant drifting issues, especially with dynamic body movements. Fortunately, due to the growing demand for sports analysis and Video Assistant Referee (VAR) systems, premier soccer events are often equipped with multi-view static camera systems, making it viable to apply multi-view markerless optical-based methods.

Hence, we leverage the capture infrastructure deployed in the 2022 FIFA World Cup stadiums to create *WorldPose*, a large-scale multi-person pose estimation dataset with accurate calibration of a moving broadcasting camera. To obtain the desired annotations from such a premier camera setup, we base our method on classic optical-based methods due to their proven robustness. However, achieving the highest possible accuracy in our setting still requires significant adaptation of said methods. This is because: 1) The distance between the cameras and the subjects is large (to ensure comprehensive coverage of the field) and movement in soccer games is fast-paced with frequent occlusions, which results in even state-of-the-art (SOTA) models experiencing a notable decline in accuracy. 2) Calibrating the moving cameras (in our case the broadcasting cameras) remains a challenge due to the rapid movement of the camera and the limited distinctive features on soccer pitches.

To address these challenges, we carefully design a data acquisition pipeline which can be summarized in the following 3 steps:

Static Camera Calibration We first calibrate the static cameras by initially treating the soccer field as a planar surface to compute the 2D homography between the image and the field plane. Then, we use the obtained homography as an initialization to solve a non-linear optimization that determines the camera parameters (including lens distortion) and accounts for the non-planar field.

Finally, an additional photometric refinement process enhances the accuracy of the camera parameters to achieve pixel-level precision.

3D Human Pose and Shape Estimation Following static camera calibration, we then estimate the 3D pose and SMPL [38] parameters in the world coordinate frame. The process starts by detecting and tracking each player's 2D keypoints. Due to the low resolution of the players in the image, we finetune the SOTA 2D detection and keypoints estimation models and leverage domain-specific knowledge to constrain the tracking algorithms. This process also includes a thorough manual review of the 2D detections with corrections if necessary. Since the static cameras are calibrated in the preceding step, we can triangulate these 2D keypoints to obtain global 3D joint coordinates. Subsequently, we fit SMPL parameters using 3D keypoint supervision, smoothness constraints, and a shape prior loss for improved accuracy.

Broadcasting Camera Calibration We first initialize the broadcasting camera parameters in a semi-automatic manner using a commercial software. However, in practice the software requires a pre-game scan by the camera operator, which is not available to us. To compensate, we leverage the 3D poses obtained from the previous step as an additional constraint, alongside the 2D field markings extracted with the software. Incorporating these additional constraints effectively enhances the accuracy and smoothness of the broadcasting calibration.

With this pipeline at hand, we curate a large-scale dataset that contains more than 2.5 million accurate 3D human pose annotations, including global player trajectories disentangled from the camera's movement, which total a travelled distance of more than 120 km. When evaluated against Vicon [58], the data acquisition pipeline yields a remarkable average error per joint of 8 cm, measured across global coordinates in a soccer stadium.

In summary, in this paper we contribute 1) WorldPose, to the best of our knowledge the first comprehensive dataset offering large-scale multi-person 3D poses paired with calibrated moving cameras. WorldPose provides accurate 3D human pose annotations with global trajectories and accurate broadcasting camera calibrations; 2) Extensive evaluations of the accuracy of our pipeline as well as baseline results of SOTA methods when evaluated on WorldPose. Our dataset and evaluation benchmarks will be made available for research.

2 Related Work

3D Human Pose Estimation Monocular 3D human pose estimation was revolutionized with the emergence of SMPL [38, 43] and more powerful Deep Learning architectures. The dominant approach is to estimate SMPL pose and shape parameters in camera-relative coordinates with a weak-perspective camera model [3,5,24,26,30,34,36,37,51,68], whereby some works consider multi-person estimation [21,31,35,42,52,54,60,62], sometimes with a focus on larger crowds in recent years [17, 21, 54, 62]. Another line of work leverages multi-view setups for multi-person pose estimation. Numerous methods [1, 10, 11, 11, 72, 74] formulate this problem as cross-view matching and association. More recent learning-based

Table 1: Comparison to related datasets. "GlobalTraj": whether poses are captured in global coordinates. "#Frames": number of frames without counting multiple views. "#Subjects": number of subjects per frame. "#Poses": total number of poses. "Camera": S (static), M (moving), M+Z (moving + zooming). *: rendered dataset.

Dataset	In the wild	GlobalTraj	Camera	#Subjects	#Frames	#Poses
KTH [28]	✓	✓	S	2	0.8k	0.8k
Panoptic Studio [25]	✗	✓	S	1-8	594k	1.5M
H3.6M [23]	✗	✓	S	1	630k	630k
PROX [18]	✗	✓	S	1-2	88k	89k
3DPW [39]	✓	✗	M	1-2	53k	75k
EgoBody [69]	✗	✓	M	2	220k	440k
RICH [22]	✗	✓	S	1-2	83k	85k
EMDB [27]	✓	✓	M	1	105k	105k
SLOPER4D [8]	✓	✓	M	1	100k	100k
BEDLAM* [2]	✓	✓	M	1-10	380k	1M
WorldPose (Ours)	✓	✓	M+Z	>10	150k	2.5M

approaches choose to directly regress 3D human pose in 3D space [6, 57, 64, 71]. Simultaneously, there has been a growing interest in the recovery of *global* human poses and camera trajectories from a single moving camera [19,32,53,65,67]. Notably, GLAMR [67] attempts to recover global trajectories from per-frame local poses. SLAHMR [65] expands on this and considers camera motions to place humans in the scenes. Other works add scene constraints to the optimization, *e.g.*, via optical flow [53] or extract background features [19, 32]. In summary, we see a clear trend in the field towards estimating 1) 3D poses of more than a handful of people and 2) with global trajectories. However, progress is severely hampered by a lack of real, in-the-wild 3D reference data. Our dataset WorldPose fills this gap and presents a challenging new setting with multiple people acting in a coordinated way in expansive space observed from moving cameras.

3D Human Pose Datasets With WorldPose we propose a dataset for monocular multi-person 3D human pose estimation, both in camera-relative and global coordinates. We highlight key differences to existing datasets in Tab. 1 and discuss them here. While a few datasets are sourced from body-worn sensors [8, 27, 39] or synthetically [2, 75], most are acquired from multi-view camera rigs [12, 22, 23, 25, 29, 40, 50, 61, 66] like ours. However, the majority of existing datasets only show 1-2 people per image. The only datasets that contain more than a handful of subjects per image are the seminal CMU Panoptic [25] and the recent BEDLAM [2] (8 or 10 subjects per image). The former is recorded in a small lab-based setting preventing dynamic captures and the latter is rendered synthetically with uncoordinated motions and no interactions. A closely related dataset is KTH Multiview Football II [28], which also features footage of soccer games. However, the 3D portion of [28] is limited in size with only 800 time instances from 3 views and 2 players. In contrast, we provide footage with 10-20 subjects per frame on average from 150k frames. In summary, in the

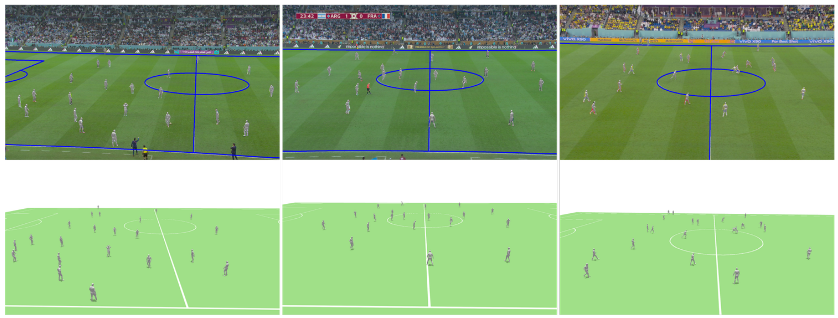

Fig. 2: Sample images of the dataset. The first row displays the camera view and overlay, and the second row presents a novel 3D view to help readers understand the 3D locations of the subjects.

landscape of 3D human pose datasets, WorldPose takes up a unique space: it contains more than double the amount of subjects per image than the previous largest dataset [2], features accurate global trajectories in a large capture area with a moving camera that is accurately calibrated, and contains high-quality SMPL pose and shape fits that are accurately tracked.

Sport Analysis Several notable works have contributed to the analysis of athletic activities in the domain of sports-related Computer Vision. [47] explores novel approaches to player and soccer scene reconstruction. [75] focuses on the reconstruction of basketball players. Both studies primarily utilize synthetic data extracted from game engines. [9, 16] provide a comprehensive soccer dataset for action analysis, albeit without 3D human poses. The domain of sports camera calibration has also been a subject of extensive study. Works such as [4, 7, 20, 44–46, 49] address challenges in achieving accurate camera calibration in dynamic sports environments. Finally, the comprehensive overview paper [56] surveys the current landscape and potential future directions at the intersection of Computer Vision and Sports Analysis. In this paper, we focus on providing a new, comprehensive dataset featuring multi-person poses aligned with a single moving camera recording professional soccer games.

3 Data

We start by describing our dataset (Sec. 3.1), capture setup (Sec. 3.2) and notations (Sec. 3.3). The data acquisition pipeline is structured into three key components (see Fig. 3): 1) calibrating static cameras around the stadium (Sec. 3.4), 2) estimating 3D human pose and SMPL parameters (Sec. 3.5), and 3) calibrating the moving broadcasting camera (Sec. 3.6).

3.1 Data Overview

We have collected a total of 88 broadcasting clips from the raw 1080p 50Hz TV program video footage of the quarter-finals and finals of the 2022 FIFA World

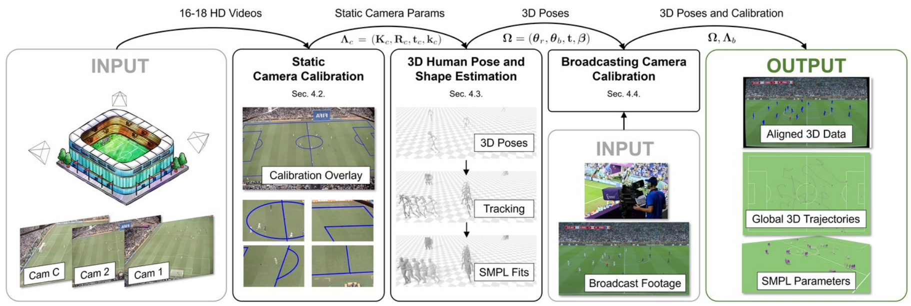

Fig. 3: Method overview (from left to right): We take as input 16-18 high-resolution videos from statically placed cameras inside the stadium. The static cameras are calibrated by using hand-picked 2D points and photometric information (Sec. 3.4). This yields camera calibrations $\boldsymbol{\Lambda}_c$ for every camera c, which we then use to triangulate and track 3D poses of each player (Sec. 3.5). We fit SMPL to the 3D pose data obtaining parameters $\boldsymbol{\Omega}$. Finally, we calibrate a moving broadcasting camera to align the estimated 3D poses with broadcasted TV footage (Sec. 3.6). The method outputs 3D SMPL pose and shape parameters $\boldsymbol{\Omega}$ of all soccer players, including their global trajectory, and accurate calibrations of the broadcast cameras $\boldsymbol{\Lambda}_b$ with high-quality player pose reprojections. Stadium image sourced from [13].

Cup. Each clip is ensured to include at least one camera pan, such that the majority of the players in action will be captured.

The resulting dataset comprises 49.7 minutes of broadcasting footage and a total of 150k frames, containing 2.5 million recorded 3D poses in SMPL format. The total global distance travelled of all subjects amounts to more than 120 km. We present visualizations of our 3D data and reprojections in Fig. 1 and Fig. 2. For additional visualization results and more statistics of the sequences, please refer to the supplementary document and video.

3.2 Capture Setup

The total capture space, equivalent to a standard soccer pitch measuring 105 $\times$ 68 meters, is covered by 16-18 strategically installed 1080p static cameras in each stadium (the number of cameras varies depending on the stadium). FIFA 3D LiDAR mappings of the World Cup stadium pitches are utilized in the static camera calibration process for better accuracy.

3.3 Notations

For each camera c of the static multi-view cameras we denote the intrinsic parameters as $\mathbf{K}_c \in \mathbb{R}^{3\times3}$, the camera distortion coefficients as $\mathbf{k}_c \in \mathbb{R}^2$, and the extrinsic parameters involving camera rotation as $\mathbf{R}_c \in SO(3)$ and translation as $\mathbf{t}_c \in \mathbb{R}^3$. We summarize all camera parameters as $\boldsymbol{\Lambda}_c = (\mathbf{K}_c, \mathbf{R}_c, \mathbf{t}_c, \mathbf{k}_c)$. Following static camera calibration, we discuss the estimation and tracking of 3D human poses $\mathbf{P} \in \mathbb{R}^{J\times 3}$ from those cameras. This process concludes with the registration of SMPL parameters, which cover shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, body pose parameters $\boldsymbol{\theta}_b \in \mathbb{R}^{69}$, root orientation $\boldsymbol{\theta}_r \in \mathbb{R}^3$, and translation $\mathbf{t} \in \mathbb{R}^3$ in world coordinates. These parameters are collectively denoted as $\boldsymbol{\Omega} = (\boldsymbol{\theta}_r, \boldsymbol{\theta}_b, \mathbf{t}, \boldsymbol{\beta})$. Finally, we model the moving broadcasting camera with frame-wise focal lengths $f \in \mathbb{R}$, principal points $\mathbf{c}_b \in \mathbb{R}^2$, radial distortion coefficients $\mathbf{k}_b \in \mathbb{R}^3$ and camera rotation $\mathbf{R}_b \in SO(3)$. We assume the camera location $\mathbf{C} \in \mathbb{R}^3$ in world space to remain constant across frames within each clip. We summarize the broadcasting camera parameters as $\boldsymbol{\Lambda}_b$.

3.4 Static Camera Calibration

For static camera calibration, we implement a multi-stage strategy inspired by the classic approach introduced in [73]. In the first stage, we approximate the soccer field as a planar surface and estimate the 2D homography between the image and the plane. Building on this, in the second stage, we utilize the homography obtained as an initialization to solve a non-linear optimization problem, aiming to determine all camera parameters, including distortion. This stage accounts for the fact that the pitch is not perfectly planar as a roughly 20 cm large field crown ensures water drainage. Finally, in the third stage, an additional photometric refinement process is applied to further refine the camera parameters, ensuring pixel-level accuracy is achieved.

Stage 1 We denote the 3D template of a soccer field as $\mathcal{S} = \{\mathbf{X}'_i\}$, which consists of real-world pitch measurements $\mathbf{X}'_i \in \mathbb{R}^3$ at characteristic field line markings obtained from official FIFA 3D LiDAR mappings of the World Cup stadium pitches. Given this template, we can estimate camera parameters by aligning the image with the projections of $\mathcal{S}$. Initially, we project the 3D markings to a flat plane, denoted as $\mathbf{x}'_i \in \mathbb{R}^2$, and manually identify 2D correspondences in the image $\mathbf{x}_{i,c} \in \mathbb{R}^2$. These 2D-to-2D correspondences are related via a 2D homography $\mathbf{x}'_i = \mathbf{H}_c\mathbf{x}_{i,c}$, whereby we can solve for $\mathbf{H}_c \in \mathbb{R}^{3\times 3}$ using the direct linear transformation algorithm. To determine the 9 parameters of $\mathbf{H}_c$ we select a few more than 4 correspondence pairs to make the problem over-determined. Through decomposition of $\mathbf{H}_c$, we obtain camera parameters $\mathbf{K}_c[\mathbf{R}_c \mid \mathbf{t}_c]$.

Stage 2 Considering that 1) the soccer pitch is not a flat plane in reality, and 2) the output of the last stage does not account for camera distortion, an additional non-linear optimization process is employed to improve the estimation. To do so, we find the closest 3D point $\mathbf{X}'_i$ for every $\mathbf{x}_{i,c}$ and then refine the camera parameters with these 3D-to-2D correspondences:

$$\boldsymbol{\Lambda}^*_c \in \arg\min_{\boldsymbol{\Lambda}_c} \sum_i \|\Pi(\mathbf{X}'_i; \boldsymbol{\Lambda}_c) - \mathbf{x}_{i,c}\|_2^2 \tag{1}$$

where $\Pi(\cdot; \boldsymbol{\Lambda}_c)$ is a non-linear function that perspectively projects 3D points into the camera c with distortion.

Stage 3 For some views, where only one or two field corners are visible, the 3D-to-2D correspondences are too sparse to provide sufficient supervision. Similar to

[47], we detect field lines in the image to obtain a denser set of correspondences. More specifically, we extract edge pixels using a line detector and construct a distance map the size of the image $\mathbf{D} \in \mathbb{R}^{H \times W}$, which for each pixel stores the distance to the nearest line pixels. We then sample new 3D points from the field lines in the 3D template $\mathcal{S}$, project them into the image and minimize the distance of the projected point to the closest line pixel via a look-up in $\mathbf{D}$:

$$\boldsymbol{\Lambda}_c^* \in \arg\min_{\boldsymbol{\Lambda}_c} \sum_{\mathbf{X}' \sim \mathcal{S}} \mathbf{D}\left[\Pi\left(\mathbf{X}'; \boldsymbol{\Lambda}_c\right)\right] \tag{2}$$

where the operator $\mathbf{D}\left[(u, v)\right]$ indexes into the matrix $\mathbf{D}$ by rounding the projected point (u, v) to integers.

3.5 3D Human Pose and Shape Estimation

With the static cameras calibrated, we now turn to estimating and tracking the 3D pose of each player, followed by fitting SMPL to the 3D pose.

Human 3D Pose Estimation and Tracking We initiate the process by detecting the bounding boxes of each player in each camera with ByteTrack [70]. Following this, we estimate the 2D poses with ViTPose [63]. As these models exhibit degraded performance on our low-resolution data, we employ a bootstrap approach to fine-tune them. Then we project the soccer field onto the images and eliminate all 2D detections located outside the field to filter out spurious detections in the audience. Given the 2D keypoints $\{\mathbf{p}_{j,c}^t \mid j \in (1, m)\}$ of the j-th player in frame t of camera c we triangulate the 3D pose denoted as $\mathbf{P}_j^t \in \mathbb{R}^{3J}$. To track a player, we associate the 2D keypoint detections with 3D pose estimations from the previous frame $\{\mathbf{P}_i^{t-1} \mid i \in (1, n)\}$ using the following affinity function A:

$$A(\mathbf{P}_i^{t-1}, \mathbf{p}_{j,c}^t) = -\operatorname{PointToRayDist}\left(\mathbf{P}_i, \Pi^{-1}(\mathbf{p}_{j,c})\right) \tag{3}$$

In other words, we compute the smallest distance of point $\mathbf{P}_i$ to the ray that results from the unprojection of $\mathbf{p}_{j,c}$ via Π^{-1}. The point-to-ray distance is averaged over the joints of the player. We do this for all i, j resulting in a $m \times n$ affinity matrix, so the tracking can be efficiently solved using a greedy matching algorithm. When 3D poses from the previous frame are not available (the first frame or when a player track is lost), we utilize epipolar distance-based association to estimate new 3D poses from unmatched 2D poses.

Bundle Adjustment Although we have achieved good alignment for the field markings after the static camera calibration stage, the presence of inevitable measurement errors has motivated us to add a bundle adjustment stage to improve the 3D pose estimation accuracy. To do so, we first hand-select a few frames per sequence where the 3D pose keypoints are of the highest quality, denoted as $\mathcal{P} = \{\mathbf{P}_j^t\}$ Assume that all the 3D player joints in $\mathcal{P}$ and 3D field markings in $\mathcal{S}$ are merged into a set $\mathcal{X}$. The bundle adjustment is then implemented by refining the camera parameters as follows:

$$\boldsymbol{\Lambda}_c^* \in \arg\min_{\boldsymbol{\Lambda}_c, \mathcal{P}} \sum_{\mathbf{X} \in \mathcal{X}} I_{j,c} \|\Pi(\mathbf{X}_j; \boldsymbol{\Lambda}_c) - \mathbf{x}_j\|^2 \tag{4}$$

where $I_{j,c}$ indicates whether point $\mathbf{X}_j$ is visible in camera c and $\mathbf{x}_j$ is the corresponding 2D point detection.

The bundle adjustment process also serves a valuable purpose by identifying outliers, where we check the points with large reprojection error after bundle adjustment and correct any mis-annotated points.

SMPL Registration and Refinement Given the 3D poses of all players, we first estimate the SMPL shape, $\boldsymbol{\beta} \in \mathbb{R}^{10}$ for each player as follows. We omit player subscripts for clarity. Assume $\mathcal{J}$ is the SMPL joint regressor, regressing 3D SMPL joints from mesh vertices. Further, let ℓ be a function that extracts all bone lengths from a skeleton into a vector of size $\mathbb{R}^{J-1}$. We can extract bone lengths from the SMPL template mesh $\bar{\mathbf{T}}$ and the shape blend shapes $\mathbf{B}_i$ and compare them to the bone lengths of $\mathbf{P}$ to directly estimate the SMPL shape $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \left| \ell(\mathbf{P}) - \ell\big(\mathcal{J}(\bar{\mathbf{T}})\big) - \sum_{i=1}^{10} \beta_i \ell\big(\mathcal{J}(\mathbf{B}_i)\big) \right| \tag{5}$$

In the next step, we align the SMPL root location to $\mathbf{P}$ by minimizing the distance between the hips and torso keypoints of the SMPL model and the corresponding joints in $\mathbf{P}$. To do so, we extract 3D keypoints from SMPL with a linear regression from the SMPL vertices, denoted as $\hat{\mathbf{P}} = \mathcal{J}(\mathcal{M}(\boldsymbol{\Omega}))$. During this optimization, we freeze all SMPL parameters except translation and global orientation.

Finally, this output serves as an initialization to fit all the SMPL parameters $\boldsymbol{\Omega} = (\boldsymbol{\theta}_r, \boldsymbol{\theta}_b, \mathbf{t}, \boldsymbol{\beta})$ with several cost terms as defined in the following. First, we employ the 2D keypoint reprojection energy term on all joints, note for optmization of poses we align the projected 2D keypoint with the hip to minimize errors introduced by a misaligned root

$$E_{\text{data}} = \|\Pi(\hat{\mathbf{P}}) - \mathbf{P}_i\|_2^2 \tag{6}$$

Additionally, we ensure smoothness of motions for each trajectory and follow [43] to incorporate shape regularization via two losses:

$$E_{\text{smooth}} = \sum_t \|(\hat{\mathbf{P}}_{t+1} - \hat{\mathbf{P}}_t) - (\hat{\mathbf{P}}_t - \hat{\mathbf{P}}_{t-1})\|_2^2 \qquad E_{\text{shape}} = \|\boldsymbol{\beta}\|_2^2 \tag{7}$$

With loss weights $\lambda_i \in \mathbb{R}_{\geq 0}$ the final loss to jointly refine all SMPL parameters is:

$$E_{\text{refine}} = \lambda_1 E_{\text{data}} + \lambda_2 E_{\text{smooth}} + \lambda_3 E_{\text{shape}} \tag{8}$$

We also observed that initializing the SMPL body poses using estimations from broadcast footage empirically improves convergence speed and performance in challenging poses.

3.6 Broadcasting Camera Calibration

A standard broadcasting camera is employed to capture live FIFA World Cup games. To calibrate it, we follow a similar strategy as for the static cameras whereby we first calibrate based on hand-picked 2D correspondences and then refine the estimation using 3D player information. However, because the broadcasting camera is moving and operates under various levels of zoom, accurate calibration is a more challenging task than for the static cameras. Furthermore, given the size of our dataset manually picking 2D correspondences is in practice not a viable solution. Thus, we use one of the leading commercial softwares built for semi-automatic calibration of broadcasting cameras [59] to facilitate an initial estimation. The use of the software allows us to manually pick field markings in a few frames per clip and the rest of the clip is then tracked autonomously by the software. In a next step, we then refine the calibrations with a dedicated optimization that ensures high-quality player and field marking reprojections, which we explain in the following.

We introduce two cost functions. For field markings, we devise a 2D reprojection regularizer to ensure that the projection of the field markings after refinement does not deviate significantly from their initial positions:

$$E_{\text{field}} = \sum_{\mathbf{X}' \in \mathcal{X}} \rho\big(\Pi(\mathbf{X}'; \boldsymbol{\Lambda}_b) - \mathbf{x}_{i,b}\big) \tag{9}$$

where ρ is the Geman-McClure function [14] and $\mathbf{x}_{i,b}$ the corresponding detections in the broadcasting camera. Additionally, we minimize the 2D reprojection loss between the 3D player keypoints and their corresponding 2D detections:

$$E_{\text{player}} = \sum_i \sum_j I_{i,j} \rho\big(\Pi(\hat{\mathbf{P}}_i; \boldsymbol{\Lambda}_b) - \mathbf{p}_j\big) \tag{10}$$

where $I_{i,j}$ is an indicator function that matches the 3D keypoints $\hat{\mathbf{P}}_i$ with the corresponding 2D keypoints $\mathbf{p}_j$. To calculate $I_{i,j}$ we perform a weighted bipartite matching process with the following similarity function:

$$\text{sim}(\mathbf{p}_i, \mathbf{p}_j) = \text{sim}_{\text{IoU}}(\mathbf{p}_i, \mathbf{p}_j) \cdot \text{sim}_{\text{bone}}(\mathbf{p}_i, \mathbf{p}_j) \tag{11}$$

where for sim_{IoU}, we calculate the intersection-over-union similarity for the bounding boxes of the players:

$$\text{sim}_{\text{IoU}}(\mathbf{p}_i, \mathbf{p}_j) = \text{IoU}\left(\text{BBox}(\mathbf{p}_i), \text{BBox}(\mathbf{p}_j)\right) \tag{12}$$

and for sim_{bone}, we calculate the mean cosine similarity between all the bones:

$$\text{sim}_{\text{bone}}(\mathbf{p}_i, \mathbf{p}_j) = \frac{1}{J-1} \sum_{k=1}^{J-1} \cos(\mathbf{p}_{i,k}, \mathbf{p}_{j,k}) \tag{13}$$

With loss weights $\lambda_i \in \mathbb{R}_{\geq 0}$ the final objective function is then

$$E_{\text{calib}} = \lambda_4 E_{\text{field}} + \lambda_5 E_{\text{player}} \tag{14}$$

Implementation details are provided in the Supp. Mat.

Fig. 4: Vicon setup at night with 6 subjects playing in the penalty box. This data is used for evaluation purposes.

4 Experiments

4.1 Metrics

We report 3 variants of the Mean Per Joint Position Error: 1) **Global MPJPE (G-MPJPE)**, where we align the entire trajectories of all players between the prediction and ground-truth using a Procrustes Alignment (PA); 2) **PA-MPJPE**: reporting the MPJPE error after aligning every player for every frame in both the prediction and the ground-truth using PA. 3) For monocular multi-person global pose estimation, we additionally report the ratio between G-MPJPE and the corresponding length of the ground-truth trajectory, which we call **Per-Meter Drift**. It quantifies the deviation of the predicted trajectory from the ground-truth trajectory per meter.

4.2 Comparison with Vicon

Table 2: Evaluation of our pipeline on Vicon setup (Sec. 3.5). BA: Bundle Adjustment, $\mathcal{P}$: set of 3D player joints, $\mathcal{S}$: set of 3D field markings, +SMPL: SMPL refinement.

	G-MPJPE [mm] ↓	PA-MPJPE [mm] ↓
Base	83.5	70.8
+ BA ($\mathcal{S}$ and $\mathcal{P}$)	86.2	70.7
+ BA ($\mathcal{S}$ only)	548.4	75.4
+ BA + SMPL	**80.0**	**66.3**

To evaluate the accuracy of our pipeline, we conduct trials with a setup featuring a Vicon [58] system to provide reference poses in a manageable portion of the pitch, *i.e.*, the penalty box. Specifically, six players equipped with Vicon markers perform common motions, including dribbling, shooting the ball, and engaging in close body contacts, as depicted in Fig. 4. Additionally, 10 synchronized static cameras are deployed around the stadium and field measurements are conducted for calibration purposes.

We run the data through our pipeline, which also includes thorough manual review of the 2D detections and association with corrections if necessary. We then compare the output poses of our method to the poses supplied by the Vicon system. We observe that the system was able to achieve a very low 6.6 cm error w.r.t. PA-MPJPE and 8.0 cm error w.r.t. G-MPJPE, which includes a measure of the global trajectory error. This underscores the high accuracy we can expect from WorldPose.

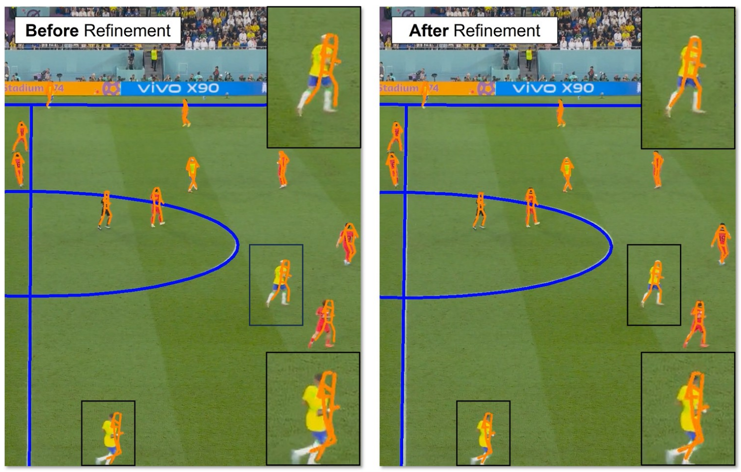

Fig. 5: Visualization of broadcasting camera calibration before (left) and after (right) refinement with 3D poses (Sec. 3.6). Note the improved reprojections in the zoom-ins.

4.3 Ablations

Multi-View Human Pose Estimation We also conducted ablation experiments to validate the design choices, as summarized in Tab. 2. It shows that the incorporation of bundle adjustment and SMPL fitting leads to improved pose estimates ("+ BA + SMPL"). Furthermore, we performed an ablation study on the bundle adjustment process, optimizing cameras with respect to field markings alone ("+ BA ($\mathcal{S}$ only)") and both field markings and keypoints ("+ BA ($\mathcal{S}$ and $\mathcal{P}$)"). The results presented in Tab. 2 indicate that while both cases significantly reduce the reprojection error for field markings, the former tends to overfit to them, resulting in worse trajectory and pose estimations.
Camera Calibration Ablation We qualitatively compare the camera calibration for broadcasting footage before and after alignment in Fig. 5. Similar to what was observed in the bundle adjustment ablation study, a low reprojection error for field markings does not necessarily imply a low reprojection error for the players. In both subfigures, reprojection for field markings is highly accurate, but misaligned player reprojections are evident without our refinement.

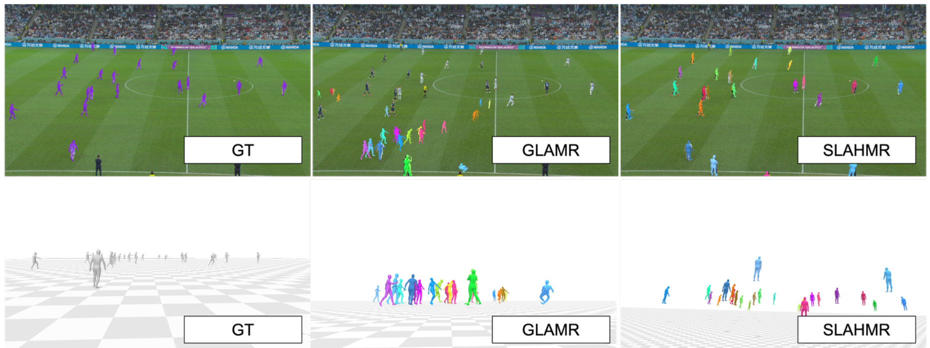

Fig. 6: Visualization of global predictions from GLAMR and SLAHMR. The first row shows the camera view and the second shows the side view.

5 Benchmarks

Table 3: Results of state-of-the-art methods on WorldPose. For "per-person" we estimate the optimal transformation for the trajectory of each player, otherwise we estimate a shared transformation for all the players.

	G-MPJPE [mm] ↓	PA-MPJPE [mm] ↓	Per-Meter Drift [cm/m] ↓
Hybrik [36]	N/A	78.8	N/A
4DHuman [17]	N/A	116.5	N/A
GLAMR [67]	18 888.9	85.2	53.3
SLAHMR [65]	8 334.1	163.9	17.6
SLAHMR w/ GT Cameras	5 837.2	199.6	10.7
SLAHMR w/o HuMoR [48]	9 736.8	140.9	20.0
GLAMR (per-person)	3 749.7	85.2	8.3
SLAHMR (per-person)	4 699.5	163.9	8.9
SLAHMR (per-person) w/ GT Cameras	3 818.5	199.6	7.2

We evaluate SOTA methods, GLAMR and SLAHMR on WorldPose. For GLAMR, we optimize the entire sequence simultaneously. For SLAHMR, as per the original paper, we first run DROID-SLAM [55] over the entire video, partition the video into chunks of 100 frames each and optimize each chunk separately. Tab. 3 and Fig. 6 summarize our key results discussed in the following:

1) While GLAMR successfully aligns all subjects on the same plane, it struggles to generate reasonable global trajectories. The prediction of SLAHMR, on the other hand, is much closer to the ground-truth. However, despite us enforcing a single shared plane, SLAHMR still fails to align the subjects on that plane.
2) To pinpoint the source of error in SLAHMR, we substitute the predicted camera parameters of DROID-SLAM with ground-truth values. This adjustment reduces the G-MPJPE and Per-Meter Drift by half. This observation

indicates that DROID-SLAM encounters difficulties in accurately predicting camera parameters, which is unsurprising considering the pitch is nearly textureless and the background (the audience seats) are highly dynamic.

3) We note that both GLAMR and SLAHMR perform worse in terms of PA-MPJPE than the method that they use for initialization (HybrIK for GLAMR, 4DHuman for SLAHMR).
4) We also note that leaving out HuMoR from SLAHMR (denoted as "SLAHMR w/o HuMoR") results in better performance. We hypothesize that this happens because HuMoR is conditioned on player height w.r.t. the plane, but the plane estimation is sometimes unreliable (see Fig. 6).
5) We additionally report the "per-person" error, where we align the trajectory of each player individually with the ground-truth. Comparing this with the "non per-person" error, we observe a significant decrease in the evaluation metrics for both GLAMR and SLAHMR, regardless of whether ground-truth cameras are used. This suggests that a significant portion of G-MPJPE arises from incorrect relative positions between the players.

With our experiments we show that while the current SOTA methods achieve impressive results in single person global human pose estimation, they: 1) encounter challenges when the area of movement expands, 2) have difficulty determining the relative positions between players, even when assuming a shared ground plane, 3) experience degraded performance when camera poses from the SLAM method are less reliable due to texture-less background or changing focal length. We believe that providing a dataset featuring data in these challenging settings will facilitate exciting new research in this area.

6 Conclusion

In this paper we present WorldPose, a novel dataset that features high-quality 3D pose, shape, and global trajectory annotations of more than 10 subjects appearing simultaneously in monocular videos. With more than 2.5 Million poses, 88 total subjects, 150k frames, and 120 km travelled distance, WorldPose is a unique dataset contributing an important building block towards advancing multi-person pose estimation and motion modelling for real-world, coordinated interactions of large groups. Our evaluations have shown that existing methods for global pose estimation struggle to produce convincing results. We hope that WorldPose will contribute to advancements of future methods in the field.

Limitations and Future work A limitation of our method is its reliance on the quality of the 2D detections and the arrangement of the static cameras, and expensive manual interventions were required when a player was not adequately covered or when the image was blurry. Another limitation is that the data focuses on male events, resulting in an unequal gender representation among the participants. We aim to expand the dataset with future access to respective data.

Acknowledgement

This work was partially supported by the Swiss SERI Consolidation Grant "AI-PERCEIVE". We would like to express our gratitude to Vizrt, particularly Janick, Fabrizio, Jens, and Daniel, for their invaluable assistance in calibrating the broadcasting video. We also thank our other friends at Vizrt for their support. Computations were carried out in part on the ETH Euler Cluster.

References

1. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1669–1676 (2014)
2. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 8726–8737 (Jun 2023)
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision – ECCV 2016. pp. 561–578. Lecture Notes in Computer Science, Springer, Springer International Publishing (Oct 2016)
4. Chen, J., Little, J.J.: Sports camera calibration via synthetic data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
5. Cho, J., Youwang, K., Oh, T.H.: Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In: European Conference on Computer Vision (ECCV) (2022)
6. Choudhury, R., Kitani, K.M., Jeni, L.A.: Tempo: Efficient multi-view pose estimation, tracking, and forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14750–14760 (2023)
7. Citraro, L., Márquez-Neila, P., Savare, S., Jayaram, V., Dubout, C., Renaut, F., Hasfura, A., Ben Shitrit, H., Fua, P.: Real-time camera pose estimation for sports fields. Machine Vision and Applications **31**, 1–13 (2020)
8. Dai, Y., Lin, Y., Lin, X., Wen, C., Xu, L., Yi, H., Shen, S., Ma, Y., Wang, C.: Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 682–692 (2023)
9. Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasrollahi, K., Ghanem, B., Moeslund, T.B., Van Droogenbroeck, M.: Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4508–4519 (2021)
10. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation from multiple views. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7792–7801 (2019)
11. Dong, Z., Song, J., Chen, X., Guo, C., Hilliges, O.: Shape-aware multi-person pose estimation from multi-view images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11158–11168 (2021)

12. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7214–7223 (2020)
13. FreePik: Designed by Freepik (2023), https://www.freepik.com/, accessed November 16th, 2023
14. Geman, S., McClure, D.E.: Statistical methods for tomographic image reconstruction. Bulletin of the International Statistical Institute **52**(4), 5–21 (1987)
15. Gemeren, C.v., Poppe, R., Veltkamp, R.C.: Spatio-temporal detection of fine-grained dyadic human interactions. In: International Workshop on Human Behavior Understanding. pp. 116–133. Springer (2016)
16. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: A scalable dataset for action spotting in soccer videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1711–1721 (2018)
17. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: ICCV (2023)
18. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: International Conference on Computer Vision (Oct 2019), https://prox.is.tue.mpg.de
19. Henning, D.F., Laidlow, T., Leutenegger, S.: Bodyslam: Joint camera localisation, mapping, and human motion tracking. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 656–673. Springer Nature Switzerland, Cham (2022)
20. Homayounfar, N., Fidler, S., Urtasun, R.: Sports field localization via deep structured models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5212–5220 (2017)
21. Huang, B., Ju, J., Li, Z., Wang, Y.: Reconstructing groups of people with hypergraph relational reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14873–14883 (October 2023)
22. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 13274–13285 (Jun 2022)
23. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (Jul 2014)
24. Iqbal, U., Xie, K., Guo, Y., Kautz, J., Molchanov, P.: Kama: 3d keypoint aware body mesh articulation. 2021 International Conference on 3D Vision (3DV) pp. 689–699 (2021), https://api.semanticscholar.org/CorpusID:233423487
25. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3334–3342 (2015)
26. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society (2018)
27. Kaufmann, M., Song, J., Guo, C., Shen, K., Jiang, T., Tang, C., Zárate, J.J., Hilliges, O.: Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14632–14643 (2023)

28. Kazemi, V., Burenius, M., Azizpour, H., Sullivan, J.: Multi-view body part recognition with random forests. In: British Machine Vision Conference (2013), https://api.semanticscholar.org/CorpusID:17823362
29. Khirodkar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., Kitani, K.: Ego-humans: An ego-centric 3d multi-human benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19807–19819 (Oct 2023)
30. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2020)
31. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: Proc. International Conference on Computer Vision (ICCV). pp. 11127–11137 (Oct 2021)
32. Kocabas, M., Yuan, Y., Molchanov, P., Guo, Y., Black, M.J., Hilliges, O., Kautz, J., Iqbal, U.: Pace: Human and motion estimation from in-the-wild videos. In: 3DV (2024)
33. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2252–2261 (2019)
34. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6050–6059 (2017)
35. Li, H., Dong, H., Jia, H., Huang, D., Kampffmeyer, M.C., Lin, L., Liang, X.: Coordinate transformer: Achieving single-stage multi-person mesh recovery from videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8744–8753 (October 2023)
36. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3383–3393 (2021)
37. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: ECCV (2022)
38. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 248 (2015)
39. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (sep 2018)
40. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3D Vision (3DV), 2017 Fifth International Conference on. IEEE (2017). https://doi.org/10.1109/3dv.2017.00064, http://gvv.mpi-inf.mpg.de/3dhp_dataset
41. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: 2018 International Conference on 3D Vision (3DV). pp. 120–130. IEEE (2018)
42. Park, S., You, E., Lee, I., Lee, J.: Towards robust and smooth 3d multi-person pose estimation from monocular videos in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14772–14782 (October 2023)

43. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10975–10985 (2019)
44. Puwein, J.: Camera Calibration and Human Pose Estimation for Sports Broadcasts and Human Performances. Ph.D. thesis, ETH Zurich (2014)
45. Puwein, J., Ziegler, R., Ballan, L., Pollefeys, M.: Ptz camera network calibration from moving people in sports broadcasts. In: 2012 IEEE Workshop on the Applications of Computer Vision (WACV). pp. 25–32. IEEE (2012)
46. Puwein, J., Ziegler, R., Vogel, J., Pollefeys, M.: Robust multi-view camera calibration for wide-baseline camera networks. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV). pp. 321–328. IEEE (2011)
47. Rematas, K., Kemelmacher-Shlizerman, I., Curless, B., Seitz, S.: Soccer on your tabletop. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4738–4747 (2018)
48. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: International Conference on Computer Vision (ICCV) (2021)
49. Sha, L., Hobbs, J., Felsen, P., Wei, X., Lucey, P., Ganguly, S.: End-to-end camera calibration for broadcast videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13627–13636 (2020)
50. Sigal, L., Balan, A., Black, M.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal on Computer Vision (IJCV) **87**(1), 4–27 (2010)
51. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 744–760. Springer (2020)
52. Sun, Y., Bao, Q., Liu, W., Fu, Y., Michael J., B., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: ICCV (2021)
53. Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M.J.: Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8856–8866 (2023). https://doi.org/10.1109/CVPR52729.2023.00855
54. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting people in their place: Monocular regression of 3d people in depth. In: CVPR (2022)
55. Teed, Z., Deng, J.: DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. Advances in neural information processing systems (2021)
56. Thomas, G., Gade, R., Moeslund, T.B., Carr, P., Hilton, A.: Computer vision for sports: Current applications and research topics. Computer Vision and Image Understanding **159**, 3–18 (2017)
57. Tu, H., Wang, C., Zeng, W.: Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 197–212. Springer (2020)
58. Vicon: Vicon Motion Systems Ltd UK (2022), https://www.vicon.com/, accessed April 10th
59. Vizrt: Viz Arena, AR Graphics and Virtual Advertising (2023), https://www.vizrt.com/products/viz-arena/, accessed November 16th, 2023
60. Wei, W.L., Lin, J.C., Liu, T.L., Liao, H.Y.M.: Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video.

In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13201–13210 (2022). https://doi.org/10.1109/CVPR52688.2022.01286
61. Wen, G., Xiaoyu, B., Xavier, A.P., Francesc, M.N.: Multi-person extreme motion prediction. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
62. Wen, H., Huang, J., Cui, H., Lin, H., Lai, Y.K., Fang, L., Li, K.: Crowd3d: Towards hundreds of people reconstruction from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8937–8946 (June 2023)
63. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (2022)
64. Ye, H., Zhu, W., Wang, C., Wu, R., Wang, Y.: Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In: European Conference on Computer Vision. pp. 142–159. Springer (2022)
65. Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling human and camera motion from videos in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21222–21232 (2023)
66. Yin, Y., Guo, C., Kaufmann, M., Zarate, J., Song, J., Hilliges, O.: Hi4d: 4d instance segmentation of close human interaction. In: Computer Vision and Pattern Recognition (CVPR) (2023)
67. Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
68. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE International Conference on Computer Vision (2021)
69. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In: European conference on computer vision (ECCV) (Oct 2022)
70. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: European Conference on Computer Vision. pp. 1–21. Springer (2022)
71. Zhang, Y., Wang, C., Wang, X., Liu, W., Zeng, W.: Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2613–2626 (2022)
72. Zhang, Y., An, L., Yu, T., Li, X., Li, K., Liu, Y.: 4d association graph for realtime multi-person motion capture using multiple video cameras. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1324–1333 (2020)
73. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on pattern analysis and machine intelligence **22**(11), 1330–1334 (2000)
74. Zhou, Z., Shuai, Q., Wang, Y., Fang, Q., Ji, X., Li, F., Bao, H., Zhou, X.: Quickpose: Real-time multi-view multi-person pose estimation in crowded scenes. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–9 (2022)
75. Zhu, L., Rematas, K., Curless, B., Seitz, S.M., Kemelmacher-Shlizerman, I.: Reconstructing nba players. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 177–194. Springer (2020)