Technical Report for Skeletal tracking 'light' challenge

Ryuto Fukushima[1*] Tianhan Xu[1*] Tatsuya Harada[1,2]

[1]The University of Tokyo [2]RIKEN AIP

{fukushima, tianhan.xu, harada}@mi.t.u-tokyo.ac.jp

Abstract

This report outlines Team mil's approach for the FIFA Skeletal Tracking Challenge 2025. The task involves accurate 3D human pose estimation from monocular soccer broadcast videos. We propose enhancements over the provided baseline, addressing its key limitations such as unreliable camera pose estimation and limited skeletal accuracy. Our method integrates optical flow-based camera pose tracking with a robust 3D human pose estimation pipeline. Ablation studies demonstrate the effectiveness of each component, and our approach achieves a significant reduction in error score on the challenge leaderboard.

1. Introduction

Estimating global 3D human pose from monocular video is a long-standing challenge in computer vision, particularly in the context of sports, where occlusion, fast motion, and limited viewpoints make the task even more complex. The FIFA Skeletal Tracking Challenge 2025 offers a practical benchmark for this task using real-world soccer broadcast footage. Participants are required to reconstruct temporally consistent 3D skeletal poses of all visible players in world coordinates, using only monocular video input and a set of known pitch landmarks.

The baseline method provided by the organizers offers a starting point, but has several limitations. It relies on moving players as visual anchors for estimating the camera pose, resulting in instability. It also uses older human pose estimation models that do not provide state-of-the-art accuracy. Furthermore, it processes each frame independently, ignoring valuable temporal continuity.

To address these limitations, we propose a robust method that combines optical flow-based camera pose estimation, a modern 3D human mesh recovery model, and novel corrections for global orientation estimation. We also introduce several enhancements, such as temporal smoothing and bounding box ensemble strategies, to improve performance in diverse scenarios.

*Equal contribution

Problem Setting Given a monocular soccer broadcast video, the goal is to estimate global 3D human poses for each visible player across all frames. The task is formulated as a sequence-to-sequence mapping, where camera intrinsics, extrinsics (provided only for the first frame), and 2D bounding boxes are given as input, and the system outputs 3D joint locations in world coordinates.

Input

- **Camera Parameters**: Intrinsic matrix (per frame), Distortion coefficients (per frame), Rotation matrix (first frame), Translation vector (first frame)
- **2D bounding boxes**: $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ for each player in the camera frame

Output

- **3D joint coordinates**: (x, y, z) for 15 selected SMPL joints per subject per frame. (nose, right / left shoulder, elbow, wrist, hip, knee, ankle, and foot.)

Evaluation Submissions are evaluated using a weighted combination of global and local Mean Per Joint Position Error (MPJPE):

$$\text{final\_score} = \text{Global MPJPE} + 10 \times \text{Local MPJPE} \quad (1)$$

- **Global MPJPE**: Measures the absolute joint error in world coordinates.
- **Local MPJPE**: Measures joint error relative to the root joint (local accuracy).

2. Baseline Method

The official baseline for the FIFA Skeletal Tracking Challenge follows a multi-stage pipeline that combines 2D/3D pose estimation with geometric reasoning to recover global 3D poses from monocular input.

The baseline consists of the following key steps:

1. **Local 2D and 3D Pose Estimation**
   For each cropped player image (based on the 2D bounding box), the 4D-Humans model [1] is applied to estimate both 2D keypoints and 3D skeletal poses

in the camera coordinate system. The output joints are reduced to 15 keypoints to match the target evaluation format.

2. **Camera Pose Estimation**
   To transform the locally estimated 3D poses into global coordinates, the following geometric steps are performed:
   (a) *Initial Camera Pose Estimation:* Players with stable aspect ratios across frames are selected, and their trajectories are used to estimate the initial camera rotation R and translation t via Singular Value Decomposition (SVD).
   (b) *Refinement Using Field Landmarks:* Known 3D pitch landmarks are first projected into the image using the initial camera estimation, and the camera poses are then refined by minimizing the reprojection error between the projected landmarks and the detected pitch lines.
3. **Global Pose Optimization**
   Once the camera parameters are determined, the local 3D poses are converted to global coordinates through the following steps:
   (a) *Translation Alignment:* Each 3D skeleton is grounded at the intersection between the ground plane (i.e., $z = 0$) and the camera ray passing through the bottom-right pixel of its 2D bounding box.
   (b) *Reprojection-based Optimization:* The global 3D pose is reprojected back into the image, and joint positions are refined by minimizing the reprojection error with respect to the detected 2D keypoints.

3. Our solution

Our solution builds on insights from the baseline method. We first identify key areas for improvement and then introduce our enhancements in detail.

- **Robust camera pose estimation:** The baseline relies on moving players as visual anchors, which can lead to unstable camera pose estimation under occlusions or abrupt movements. We propose an alternative camera pose estimation method based on pitch landmarks using optical flow.
- **Modern pose estimation backbone:** The use of 4D-Humans[1], while effective, does not leverage recent advancements in mesh-based human pose estimation models. We adopt SMPLest-X[4] for higher accuracy and richer joint representation.
- **Global orientation compensation:** Most human mesh recovery models assume a weak perspective projection, which can lead to incorrect global orientations – especially for people near the image boundaries. We explicitly correct the predicted global orientation by considering the angle between the camera's optical axis and the viewing ray of each bounding box, resulting in more accurate and stable orientation estimates.
- **Temporal consistency:** The baseline treats each frame independently. Our method incorporates temporal smoothing and filtering strategies to ensure stable 3D trajectories across frames.

Figure 1. Tracking of pitch landmarks using optical flow. Red: projected positions (current frame); Blue: predicted positions (next frame).

3.1. Camera Pose Estimation

We begin by projecting 3D pitch landmarks into the image frame using known camera parameters. For the next frame, we predict the optical flow of these landmarks to obtain correspondences.

Using the matched points from adjacent frames, we apply the Perspective-n-Point (PnP) algorithm to estimate the relative camera pose. This procedure is repeated frame-by-frame to estimate the camera pose trajectory throughout the sequence. Figure 2 shows a visual comparison of baseline vs. our method. While the baseline often fails to maintain stable estimates over long sequences, our method remains robust even under extended and challenging conditions

3.2. 3D Skeletal Pose Estimation

We replace the baseline's 3D pose estimator with SMPLest-X [4], which offers significantly improved accuracy and expressiveness. To correct errors arising from the weak perspective assumption, we adjust the predicted global orientation based on the ray direction between the bounding box center and the camera's optical axis (Figure 3). As shown in Figure 4, the compensated skeleton better matches the 2D image, particularly at the spine alignment.

3.3. Additional Enhancements

In addition to the core improvements, we apply several minor yet effective techniques to further enhance pose estimation:

Figure 2. Projected pitch points (frame 300). Top: Baseline. Bottom: Ours.

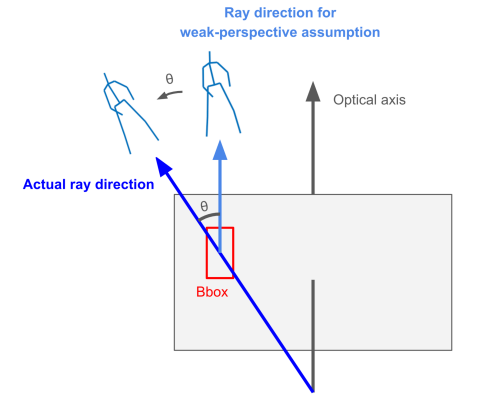

Figure 3. Illustration of global orientation correction based on ray-angle with camera optical axis.

- **Tangent space optimization:** Body rotations are optimized in tangent space, resulting in smoother and more stable orientation estimates compared to direct parameter updates.
- **Temporal smoothing:** Gaussian filtering is applied to joint trajectories over time, improving temporal consistency and reducing jitter.
- **Ensemble strategies:** We generate multiple predictions by varying the size of input bounding boxes and aggregate them. This approach improves robustness, particularly under detection noise or occlusion.
- **Pseudo pitch landmarks:** We introduce additional virtual pitch landmarks in areas with sparse field features, such as between the center circle and penalty box.

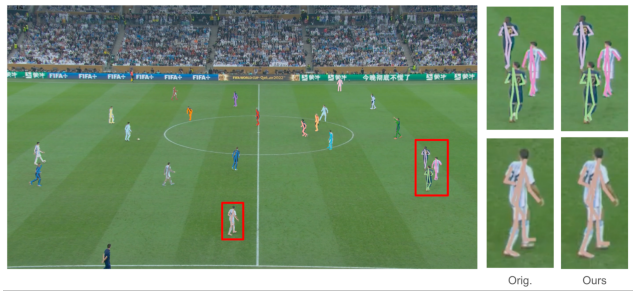

Figure 4. Qualitative comparison of global orientation with (right) and without (left) compensation.

3.4. Ablation Study

To evaluate the contribution of each component in our pipeline, we conducted an ablation study using the official leaderboard metric (lower is better). Table 1 summarizes the step-by-step improvements.

Table 1. Ablation results showing the leaderboard score after each successive improvement.

Method Variant	Leaderboard Score
Baseline (4D-Humans, no camera fix)	6.23
+ Optical flow & PnP for camera pose	1.85
+ Tangent space optimization	1.71
+ Temporal smoothing (hip joint only)	1.68
+ Replace with SMPLest-X	1.53
+ Temporal smoothing (all joints)	1.48
+ Ensemble bounding boxes	1.41
+ Global orientation compensation	**1.37**

The results highlight the cumulative impact of each enhancement. In particular, our robust camera pose estimation contributes substantially to the overall performance gain.

4. Discussion and Future Work

Our method achieves substantial error reduction by improving both camera pose estimation and 3D human pose representation. However, several directions remain open for further enhancement:

- **Learning-based Optical Flow:** While we currently use a classical Lucas-Kanade optical flow method for simplicity, future work could explore learning-based approaches

such as RAFT [3], which may offer improved landmark tracking performance.

- **Multi-frame Pose Estimation:** Incorporating temporal context directly into the pose estimation model may improve consistency without post-smoothing.
- **Domain-specific Fine-tuning:** Training the pose estimation model on the WorldPose dataset [2] may lead to improved accuracy in soccer-specific scenarios.

5. Acknowledgment

This research is partially supported by JST Moonshot R&D Grant Number JPMJPS2011, CREST Grant Number JPMJCR2015 and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo.

References

[1] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[2] Tianjian Jiang, Johsan Billingham, Sebastian Müksch, Juan Zarate, Nicolas Evans, Martin R. Oswald, Marc Pollefeys, Otmar Hilliges, Manuel Kaufmann, and Jie Song. Worldpose: A world cup dataset for global 3d human pose estimation, 2025. 4

[3] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. 4

[4] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. Smplest-x: Ultimate scaling for expressive human pose and shape estimation, 2025. 2