2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Low-latency online estimation of human upper-limb pose and kinematics from a single 360° camera

Mathis D'Haene Guillaume Caron Yusuke Yoshiyasu Bruno Watier

Abstract— **We present a fully online framework for streaming human upper-limb kinematics estimation from a single 360° camera. Incoming frames are processed sequentially through vertical-boundary-aware tracking, pseudo-perspective rendering, and Neural Localizer Fields to estimate a sparse set of 3D anatomical landmarks in real time. These landmarks are mapped to an OpenSim-compatible biomechanical model, with joint angles computed on the fly via an online inverse kinematics solver. The system achieves end-to-end latencies as low as 22.9 ms on a high-performance setup. Evaluated in a single-participant scenario involving an initial T-pose calibration and repeated object displacement toward the camera, it demonstrates robust performance under moderate self-occlusion and spherical distortion. While tested in a constrained setting, its modular, real-time design makes it a promising candidate for human–robot interaction and other motion analysis applications, enabling minimal, markerless, and anatomically interpretable upper-limb tracking from omnidirectional vision.**

Index Terms— Markerless Motion Capture, Omnidirectional Imaging, Human–Robot Interaction, Real-Time Tracking, Neural Localizer Fields, Biomechanics, Human Movement Analysis

I. Introduction

In human–robot interaction, accurate perception of human motion is crucial for enabling safe and responsive collaboration in dynamic and unstructured environments [1]. While many tasks can benefit from detailed motion understanding, upper-limb kinematics are particularly important for anticipating and adapting to human actions. However, robots often fail to fully interpret human intentions due to limited real-time perception of anatomically meaningful upper-limb movements [2].

Mathis D'Haene is with the Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS), Centre National de la Recherche Scientifique (CNRS), Université de Toulouse—UPS, 31000 Toulouse, France, and with the CNRS–AIST JRL (IRL), National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba 305-8560, Ibaraki, Japan. mathis.d-haene@laas.fr.

Guillaume Caron is with the CNRS–AIST JRL (IRL), National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba 305-8560, Ibaraki, Japan, and with the MIS Laboratory, Université de Picardie Jules Verne, Amiens, France. guillaume.caron@u-picardie.fr.

Yusuke Yoshiyasu is with the Computer Vision Research Team, Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, 305-8560, Japan, and with the CNRS–AIST JRL (IRL), AIST, Tsukuba 305-8560, Ibaraki, Japan. yusuke.yoshiyasu@aist.go.jp.

Bruno Watier is with the Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS), Centre National de la Recherche Scientifique (CNRS), Université de Toulouse—UPS, 31000 Toulouse, France. bruno.watier@laas.fr.

* This work is supported by the ANR – FRANCE (French National Research Agency) under the CALL project n° ANR-24-CE10-5850-01.

Traditionally, precise motion capture systems, such as optical Vicon or Optitrack commercial systems, have been the gold standard for extracting kinematics, but they are expensive, intrusive, and impractical in real-world settings. In contrast, recent advances in Human Pose and Shape Estimation (HPSE) from monocular images or videos have shown promising results [3]–[5]. However, most methods rely on standard camera perspectives [6] and lack anatomically meaningful keypoints [7], [8]. In addition, most human pose and shape estimation models are too computationally demanding for deployment in real-time robotic applications [9].

360° cameras offer a unique advantage for robot-mounted perception by capturing the full spherical environment without blind spots [10]. Yet their use in human pose and shape estimation remains limited. Compared to perspective or fisheye views, omnidirectional images face unique challenges: distortions across the entire frame, scale changes with position, and subject duplication when crossing the vertical boundaries (where the left and right edges of the equirectangular projection meet). These effects reduce 2D detection accuracy and disrupt temporal tracking. Consequently, most prior work has focused on depth estimation, segmentation, or large-scene object detection [11], [12], while human motion capture efforts have typically used fisheye cameras and targeted egocentric or ambient scene analysis rather than close-range pose reconstruction [13]–[15].

In this work, we aim to design and evaluate a real-time, online pipeline for estimating anatomically informed upper-limb joint angles from omnidirectional video. Here, the term upper limb refers to the right arm together with the torso segment required for joint angle computation. The system integrates pseudo-perspective rendering, vertical-boundary-aware tracking, Neural Localizer Fields (NLF), and OpenSim-compatible inverse kinematics into a sequential, low-latency processing chain.

To assess its feasibility, we evaluate the pipeline in a controlled single-participant scenario involving an initial T-pose calibration followed by repeated object displacement tasks toward the camera, simulating interaction with a robot equipped with a fixed 360° viewpoint. This setup introduces moderate self-occlusion, and natural motion dynamics, while enabling controlled measurement of processing latency.

We seek to demonstrate that a single 360° camera can provide sufficient visual information to generate anatomically consistent and temporally smooth kinematics in such a setting. The modular design is intended to facilitate adaptation to different hardware configurations and application contexts, while maintaining interpretable outputs compatible with

biomechanical analysis. We anticipate that this perception-only approach can operate in real time on modern GPUs, paving the way for integration into responsive human–robot interaction systems.

II. Related Work

A. Human Pose-and-Shape Estimation

Human pose and shape estimation has seen major progress in the past decade. Early methods relied on sparse keypoint detection using stickman-style representations [16], while more recent approaches estimate full-body mesh parameters using parametric models such as the Skinned Multi-Person Linear Model (SMPL) [17]. Image-based pipelines like HMR2.0 [9], SAT-HMR [18], and CameraHMR [6] offer single-frame inference, whereas video-based models such as VIBE [19] leverage temporal context to improve motion consistency, though at the cost of higher latency. In parallel, non-parametric methods—which avoid regressing SMPL parameters—have gained traction for their flexibility and efficiency [20]. Neural Localizer Fields (NLF) [21], in particular, directly regress the 3D positions of arbitrary mesh vertices or keypoints from RGB input, without relying on a global pose or shape representation.

B. 360° Vision for Human Pose and Shape Estimation

Modern 360° cameras often capture the surrounding environment using dual fisheye lenses, which are then stitched together into a single 2D equirectangular projection. In this format, latitude and longitude angles are unwrapped into a rectangular image. While convenient for processing, this projection introduces distortions—particularly near the poles and along the vertical boundary—that require careful handling [22]. While some studies use egocentric fisheye cameras for environmental awareness and body tracking [14], the use of 360° cameras for pose estimation is still limited. Most systems rely on undistortion or pseudo-perspective transformations [13], [23]. However, to date, no prior work has attempted real-time human tracking, pose estimation, and biomechanical inference from 360° video in contexts involving close-range object manipulation, including but not limited to human–robot interaction.

C. Biomechanics and Human Pose-and-Shape Estimation

Biomechanical analysis from image data typically requires anatomical joint angles, which are difficult to extract from simple keypoints from stickman-like models [16]. Marker augmentation methods [7] estimate joint angles from 3D stickman-like sequences but discard image content and struggle to generalize to unseen motions, requiring extensive and diverse datasets for reliability across human movements. SMPL-based mesh recovery offers realistic body shape, but the joint angles are not biomechanically consistent [24]. Keller et al. [24] introduced SKEL, a pipeline that derives OpenSim-based joint angles from SMPL meshes through optimization. However, its high computational cost prevents real-time inference.

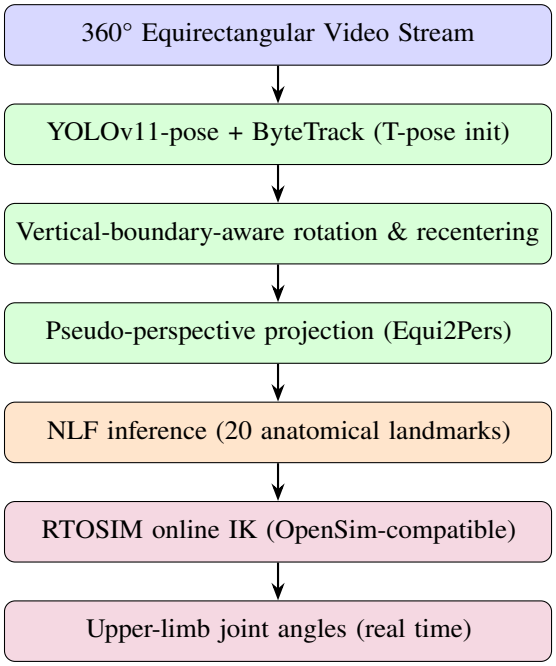

Fig. 1. Sequential, online pipeline from a 360° equirectangular video stream to upper-limb joint angles. Color coding indicates the processing stage: **blue** – video acquisition from the camera, **green** – pre-processing to prepare the image for inference, **orange** – estimation of anatomical landmarks, **purple** – OpenSim-compatible real-time inverse kinematics and joint angle computation.

D. Understanding Upper-Limb Motion

Accurately characterizing upper-limb kinematics is essential in many human–robot interaction scenarios, from collaborative assembly to gesture-based control. However, most prior work has addressed other scales or focus areas — either detailed hand modeling [25], or full-body pose recovery in human–object interactions [26] — without specifically targeting anatomically meaningful upper-limb parameters. A notable exception is the work of Xiang et al. [27], who estimated joint angles from multi-camera RGB input in a markerless setup. Nevertheless, their method relies on sparse keypoints—essentially a stick-figure representation — which cannot recover important biomechanical degrees of freedom.

III. Method

We propose a modular pipeline, summarized in Figure 1, composed of the following stages: detection and tracking, pseudo-perspective rendering, 3D pose estimation, and online inverse kinematics. While each stage is implemented as an independent module, they operate sequentially on the same incoming frame, enabling fully online processing without future context. This design supports low-latency performance and allows straightforward integration into reactive systems.

- **360° live streaming:** The Ricoh Theta X [28], a dual-fisheye 360° camera, captures the entire surrounding environment, as illustrated in Figure 2. For live streaming, the camera is connected via USB and exposed as a USB Video Class peripheral. The video stream is configured at a resolution of 3840×1920 pixels, encoded

in H.264 at 29.97 frames per second, providing an efficient equirectangular representation with a good trade-off between compression, latency, and image quality. Frame acquisition is performed with a modified version of the libuvc python library [29], providing low-level access to the Ricoh Theta X stream and integrating with GStreamer. The retrieved frame (in RGB format) is then forwarded to the detection and tracking pipeline.

- **Detection and Tracking:** We employ YOLOv11-pose [30] for human detection and keypoint estimation, combined with ByteTrack [31] for temporal tracking. Tracking initialization is performed during a three-second T-pose, allowing the assignment of a unique ID to the target subject (Fig. 3). In equirectangular images, the subject may appear split across the two vertical edges (one part at the right edge, the other at the left). Following [32], we perform detection on both the original and a 180°-rotated view, discard boxes near image boundaries, and track the most centered view. At each frame, the view is dynamically rotated to keep the subject away from the vertical boundaries, preventing ID changes when crossing them (Fig. 4).
- **Pseudo-perspective projection:** We apply a GPU-accelerated version of Equi2Pers from the EquiLib Python library [33] to generate a pseudo-perspective view—i.e., a perspective-like crop from the 360° equirectangular input (Fig. 5). This transformation maps a narrow field of view (55°) onto a rectilinear image centered on the tracked subject. The virtual camera is centered using the bounding box geometry: we take the horizontal midpoint of the box and a vertical point located at two-thirds of its height from the bottom. This ensures that the view remains focused on the upper body regardless of subject orientation. The resulting image has a resolution of 256×256 pixels, matching the input resolution and field of view used to train the subsequent pose estimation model, and preserves local geometric consistency suitable for 2D inference models.
- **Keypoints extraction:** We start from a standard SMPL mesh, where each vertex corresponds to a fixed 3D surface location. In this mesh, each vertex corresponds to a fixed 3D location on the surface model. Before running the pipeline, we manually defined 20 vertices corresponding to standardized anatomical landmarks [8], covering the right upper limb. This selection was performed offline in Blender and remains fixed across all experiments. The chosen vertices align with a biomechanical model of the upper limb featuring 13 degrees of freedom (six for the trunk, three for the shoulder, two for the elbow, and two for the wrist) [34]. At runtime, the pipeline directly predicts the 3D positions of these vertices, which are then mapped to an OpenSim-compatible biomechanical model for joint angle computation. We use the NLF-S variant of Neural Localizer Fields (NLF) [21], a non-parametric human pose-and-shape estimation model, to directly predict the 3D coordinates of these anatomical landmarks. Unlike larger NLF models, NLF-S offers a favorable trade-off between runtime and accuracy, making it suitable for real-time use. Instead of regressing the full mesh or estimating SMPL pose parameters, NLF infers the positions of specific mesh vertices given a single RGB image. This approach preserves anatomical interpretability while keeping computational cost low. The selected vertex IDs are visualized in Figure 5, and form the basis for the downstream biomechanical analysis.
- **Online kinematics analysis:** We use RTOSIM, an open-source inverse kinematics library built on OpenSim [35], to compute joint angles consistent with the recommendations of the International Society of Biomechanics [8]. The anatomical landmarks predicted by NLF are transmitted to RTOSIM through a socket-based interface. Figure 5 shows the resulting biomechanical model performing an object displacement task. Finally, the estimated angles are smoothed using a causal 4th-order Butterworth low-pass filter with a 6 Hz cutoff, reducing jitter while preserving motion dynamics, as illustrated in Figure 7. At a 30 Hz sampling rate, this configuration introduces an effective passband delay of approximately 70 ms.

IV. Results

A. Experimental Protocol

The evaluation was conducted in a dynamic scenario designed to reflect realistic operation rather than static imagery. A single participant performed an initial T-pose calibration, followed by three repetitions of the same object-displacement task, in which an object held in the right hand was brought toward the camera and then lowered, simulating interaction with a robot equipped with a 360° camera on a fixed base. This protocol was chosen to ensure that runtime measurements captured the effects of natural motion, and moderate self-occlusion.

B. Performance and Qualitative Results

Using the above protocol, Table I summarizes the latency of each component of the pipeline across two hardware setups. On the standard GPU-equipped system (Intel Ultra 5 125U + NVIDIA A30), the full pipeline operates with an end-to-end latency of 72.9 ms per frame (13.7 fps). On the high-performance setup (Intel i9-13900K + NVIDIA RTX 4090), latency drops to 22.9 ms per frame (43.6 fps).

The most time-consuming step on both setups is the NLF inference, followed by detection and tracking. Input latency from GStreamer also varies greatly between systems, primarily due to differences in CPU performance. On high-performance setup, the pipeline can process frames at nearly 44 fps. However, the Ricoh Theta X camera outputs at a maximum of 29.97 fps at the tested resolution, which sets an upper bound on the achievable real-time throughput regardless of computational speed.

Importantly, the modular structure of the system enables further speedups through parallelization. On the A30 GPU, decoupling image acquisition and inference into separate

TABLE I

LATENCY (MS) COMPARISON ACROSS HARDWARE SETUPS (MEAN ± STD)

Component	Standard GPU (A30)	High-perf. (RTX 4090)
YOLO + Track + Rot.	17.6 ± 1.2	8.1 ± 0.1
Equi2Pers (GPU)	2.9 ± 0.0	1.1 ± 0.0
NLF Inference	21.4 ± 0.3	7.5 ± 0.1
RTOSIM Solver	1.0 ± 0.2	0.2 ± 0.0
Subtotal (inc. overhead)	**42.9 ± 1.4**	**17.9 ± 0.2**
GStreamer Input	30.0 ± 1.1	5.0 ± 0.3
Total (inc. overhead)	**72.9 ± 1.8**	**22.9 ± 0.4**

Std. GPU: Intel Ultra 5 125U + NVIDIA A30
High-perf.: Intel i9-13900K + NVIDIA RTX 4090

threads could increase throughput toward the NLF step's bottleneck of 21.4 ms per frame. Fully parallelizing all blocks could further reduce waiting time between stages. On the RTX 4090 GPU, the overall processing speed already exceeds the camera's 29.97 fps output limit, so parallelization would mainly reduce latency rather than increase throughput.

Figures 2–4 show intermediate results from the early pipeline stages: T-pose detection (Fig.2), robust tracking (Fig.3), and dynamic view recentering (Fig.4). Figure 5 illustrates the final stages for one of the three object displacements performed during the evaluation, including virtual camera generation, NLF-based landmark inference, and biomechanical modeling. The pipeline outputs a clean pseudo-perspective view, temporally consistent landmarks, and real-time joint angle estimates—demonstrating coherent operation across modules.

We also provide a supplementary video that is longer than the evaluation extract and includes additional viewpoints and multiple object displacement repetitions. The experimental evaluation presented here is based on a 10-second segment from this video (seconds 18–28), which contains three object displacements. Figure 5 corresponds to one of these displacements, while Figure 7 compares markerless and motion-capture joint angles over all three repetitions. Although the curves appear well aligned, a full validation across multiple subjects remains for future work.

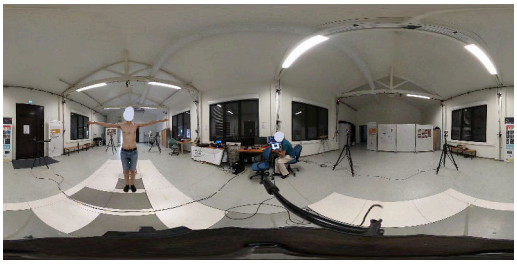

Fig. 2. Equirectangular projection of the subject performing a T-pose

V. DISCUSSION AND LIMITATIONS

This work demonstrates the feasibility of estimating anatomically consistent upper-limb kinematics from a single 360° camera in a realistic, dynamic scenario. The results confirm that all processing stages, from omnidirectional image acquisition to biomechanically compatible joint angles, can run

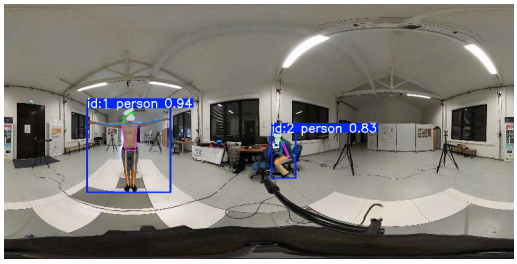

Fig. 3. The subject holding a T-pose for a certain time is tracked (using YOLOv11-pose and ByteTrack).

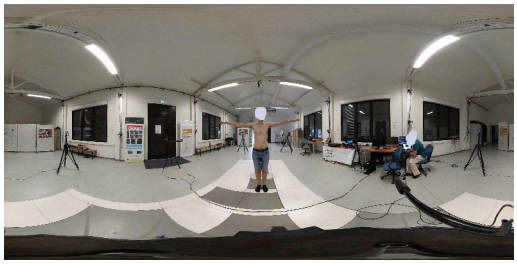

Fig. 4. The tracked subject is dynamically centered in the middle of the frame.

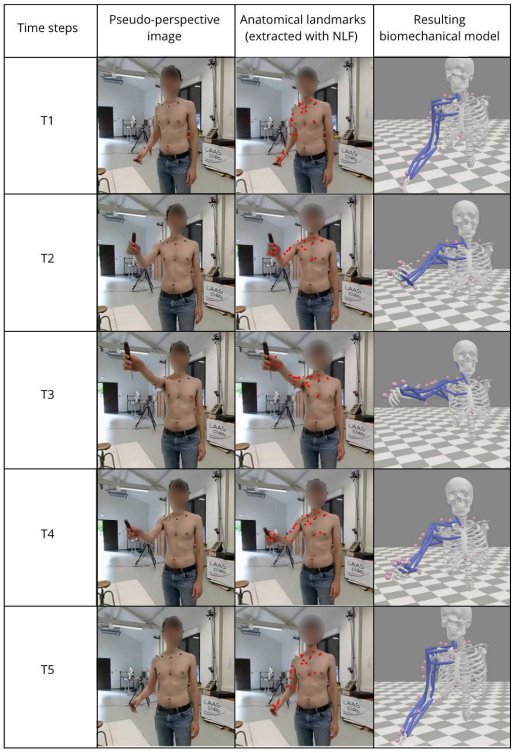

Fig. 5. Example frames (T1–T5) from an object displacement sequence, illustrating the final stages of the proposed pipeline: (left) pseudo-perspective view centered on the tracked subject, (middle) anatomical landmarks predicted by NLF (in red), and (right) resulting biomechanical model in OpenSim (see supplementary video).

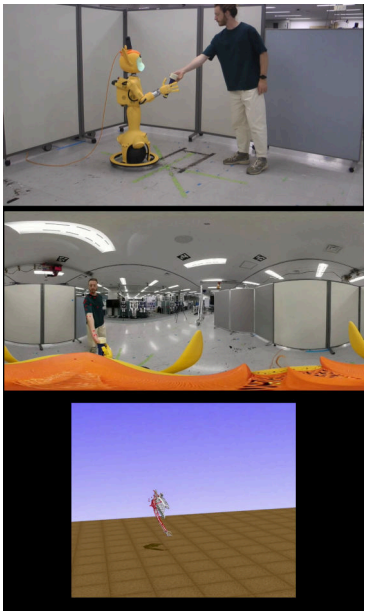

Fig. 6. Demonstration of the proposed pipeline mounted on the head of the Miroki robot (Enchanted Tools, France). Top: external view of the robot equipped with the Ricoh Theta X camera. Middle: Miroki's 360° point-of-view with anatomical landmarks predicted by NLF (in red). Bottom: corresponding OpenSim-based inverse-kinematics reconstruction of the upper-limb motion.

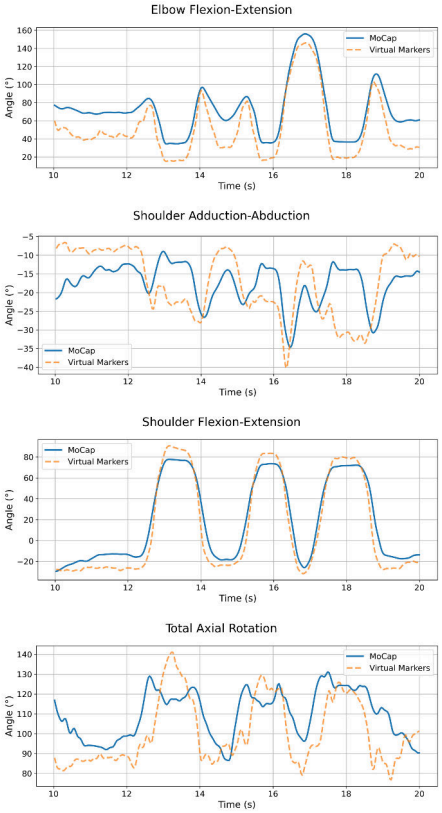

Fig. 7. Comparison of joint angles between MoCap and markerless data of a subject performing 3 object displacement tasks in a row. Four upper-limb degrees of freedom are presented.

sequentially in real time on modern GPUs, with end-to-end latencies between 22.9 ms and 75 ms depending on hardware. By integrating non-parametric mesh vertex prediction with a biomechanical model, the proposed approach leverages recent advances in markerless motion capture to meet the latency and interpretability requirements of responsive human–robot interaction.

The evaluation—an initial T-pose calibration followed by repeated object displacement tasks toward the camera—introduced moderate self-occlusion, and natural movement patterns. In this setting, the predicted landmarks and joint angles appeared plausible and temporally smooth (Fig. 7), suggesting that a single omnidirectional viewpoint can provide sufficient visual information for anatomically meaningful motion estimation under controlled conditions.

However, further work is needed to quantitatively assess the accuracy of the 3D joint angle estimates. The present comparison with motion capture ground truth remains qualitative and limited to a single participant. In particular, the anatomical fidelity of complex motions, such as arm abduction–adduction or wrist rotations, has yet to be evaluated through standardized benchmarks and multi-subject studies, and against established markerless pipelines such as OpenCap's Marker Augmenter [7] or HSMR [36].

While our system relies on only a single 360° camera, we do not claim it eliminates the need for external sensors in all contexts. Rather, it points toward minimal sensor configurations for upper-body kinematics capture, especially where multi-camera setups are impractical. The system qualifies as both *real-time* and *online* in terms of processing flow and latency. A key direction for ongoing work is integration into closed-loop human–robot interaction controllers, so that real-time kinematic feedback can directly inform motion planning and execution during object displacement tasks.

Beyond the fixed-base evaluation, we also tested the proposed system on a moving base by mounting the Ricoh Theta X on the head of Miroki — a collaborative robot developed by Enchanted Tools (Fig. 6). Although no active handover was performed, this configuration enabled us to qualitatively verify that the pipeline remains operational under moderate camera motion and background variation. These exploratory trials suggest that the perception-only design can be extended to mobile robotic platforms without major modification. The accompanying video illustrates both the fixed-base and the moving-base experiments, including the Miroki robot demonstration shown in Figure 6.

Current limitations include the single-person assumption and the need for a T-pose initialization. Multi-person tracking in spherical views remains challenging, particularly near the vertical boundaries of the equirectangular format where ID switches can occur. Additionally, proximity to the camera increases spherical distortion, degrading detection and inference reliability. While our observations concern 360° imagery, a similar non-trivial relationship between subject–camera distance and distortion effects has been reported for perspective images [37]. Addressing these issues—together with broader robustness and generalization—will be essential for

deployment in unconstrained interaction scenarios.

VI. Conclusion

We presented a modular and low-latency pipeline for estimating upper-limb 3D kinematics from a single 360° camera, designed for online operation in human–robot interaction contexts. By integrating equirectangular streaming, pseudo-perspective projection, lightweight keypoint inference with Neural Localizer Fields, and OpenSim-compatible inverse kinematics, the system achieves anatomically meaningful upper-limb pose and kinematics estimation in real time. While we do not yet model full interaction or robot control, our results demonstrate the feasibility of markerless, minimal-sensor kinematic tracking using omnidirectional vision. Future work will focus on benchmarking against ground-truth motion capture, extending to dynamic scenes and multiple subjects, and closing the loop with robotic feedback and decision-making.

References

[1] V. Ortenzi, A. Cosgun, T. Pardi, W. P. Chan, E. Croft, and D. Kulić, "Object handovers: A review for robotics," *IEEE Trans. on Robotics*, vol. 37, no. 6, pp. 1855–1873, 2021.

[2] A. Castro, F. Silva, and V. Santos, "Trends of human-robot collaboration in industry contexts: Handover, learning, and metrics," *Sensors*, vol. 21, no. 12, 2021.

[3] T. Jiang, X. Xie, and Y. Li, "RTMW: Real-Time Multi-Person 2D and 3D Whole-body Pose Estimation," *arXiv e-prints*, p. arXiv:2407.08634, Jul. 2024.

[4] Y. Huang, J. Liu, K. Xian, and R. C. Qiu, "Posemamba: Monocular 3d human pose estimation with bidirectional global-local spatio-temporal state space model," *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 39, no. 4, pp. 3842–3850, Apr. 2025.

[5] S. Mehraban, V. Adeli, and B. Taati, "Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Comput. Vis. (WACV)*, January 2024, pp. 6920–6930.

[6] P. Patel and M. J. Black, "CameraHMR: Aligning people with perspective," in *Int. Conf. on 3D Vis. (3DV)*, 2025.

[7] A. Falisse, S. D. Uhlrich, A. S. Chaudhari, J. L. Hicks, and S. L. Delp, "Marker data enhancement for markerless motion capture," *IEEE Trans. on Biomedical Engineering*, vol. 72, no. 6, pp. 2013–2022, 2025.

[8] G. Wu, F. C. van der Helm, H. (DirkJan) Veeger, M. Makhsous, P. Van Roy, C. Anglin, J. Nagels, A. R. Karduna, K. McQuade, X. Wang, F. W. Werner, and B. Buchholz, "Isb recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—part ii: shoulder, elbow, wrist and hand," *Journal of Biomechanics*, vol. 38, no. 5, pp. 981–992, 2005.

[9] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4d: Reconstructing and tracking humans with transformers," in *Proc. of the IEEE/CVF Int. Conf. on Comput. Vis. (ICCV)*, October 2023, pp. 14 783–14 794.

[10] A. N. André, F. Morbidi, and G. Caron, "Uniphorm: A new uniform spherical image representation for robotic vision," *IEEE Transactions on Robotics*, vol. 41, pp. 2322–2339, 2025.

[11] Y. Xu, H. Huang, Y. Chen, and S.-K. Yeung, "360VOTS: Visual Object Tracking and Segmentation in Omnidirectional Videos," *arXiv e-prints*, p. arXiv:2404.13953, Apr. 2024.

[12] Y. Guo, S. Garg, S. M. H. Miangoleh, X. Huang, and L. Ren, "Depth any camera: Zero-shot metric depth estimation from any camera," in *Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, June 2025, pp. 26 996–27 006.

[13] K. Deshpande, C. Heindl, G. Stübl, M. J. Kollingbaum, and A. Pichler, "Novel first person view for human 3d pose estimation in robotic applications using fisheye cameras," in *2024 10th Int. Conf. on Automation, Robotics and Applications (ICARA)*, 2024, pp. 112–116.

[14] Y. Liu, J. Yang, X. Gu, Y. Chen, Y. Guo, and G.-Z. Yang, "Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning," *IEEE Trans. on Multimedia*, vol. 25, pp. 8880–8891, 2023.

[15] Y. Zhang, S. You, S. Karaoglu, and T. Gevers, "Multi-person 3d pose estimation from a single image captured by a fisheye camera," *Comput. Vis. and Image Understanding*, vol. 222, p. 103505, 2022.

[16] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2014.

[17] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Trans. on Graph. (TOG)*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[18] C. Su, X. Ma, J. Su, and Y. Wang, "Sat-hmr: Real-time multi-person 3d mesh estimation via scale-adaptive tokens," in *Proc. of the Comput. Vis. and Pattern Recognit. Conf. (CVPR)*, June 2025, pp. 16 796–16 806.

[19] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, June 2020.

[20] Y. Yoshiyasu, "Deformable mesh transformer for 3d human mesh recovery," in *Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, June 2023, pp. 17 006–17 015.

[21] I. Sárándi and G. Pons-Moll, "Neural localizer fields for continuous 3d human pose and shape estimation," 2024.

[22] I.-C. Lo, K.-T. Shih, and H. H. Chen, "Efficient and accurate stitching for 360° dual-fisheye images and videos," *IEEE Trans. on Image Processing*, vol. 31, pp. 251–262, 2022.

[23] K. Aso, D.-H. Hwang, and H. Koike, "Portable 3d human pose estimation for human-human interaction using a chest-mounted fisheye camera," in *Proc. of the Augmented Humans Int. Conf. 2021*, ser. AHs '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 116–120.

[24] M. Keller, K. Werling, S. Shin, S. Delp, S. Pujades, C. K. Liu, and M. J. Black, "From skin to skeleton: Towards biomechanically accurate 3D digital humans," *ACM Trans. on Graph. (ToG)*, vol. 42, no. 6, pp. 253:1–253:15, Dec. 2023.

[25] H. Xu, H. Li, Y. Wang, S. Liu, and C.-W. Fu, "Handbooster: Boosting 3d hand-mesh reconstruction by conditional synthesis and sampling of hand-object interactions," in *Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, June 2024, pp. 10 159–10 169.

[26] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Behave: Dataset and method for tracking human object interactions," in *Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, June 2022, pp. 15 935–15 946.

[27] Y. Xiang, B. Yao, M. Ba, W. Xu, Y. Tang, and L. Li, "Human grasping behavior analysis in human-robot collaboration based on markerless motion capture," in *2024 IEEE 20th Int. Conf. on Automation Science and Engineering (CASE)*, 2024, pp. 534–539.

[28] Ricoh Company, Ltd., "Ricoh THETA X – 360 Camera," 2022, accessed: August 2025.

[29] Nickel110, "gstthetauvc: Streaming ricoh theta x via gstreamer and libuvc," 2023, accessed: 2025-08-04.

[30] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[31] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," 2022.

[32] A. Farid and O. Yoshie, "Monocular object detection localization on a 2d plane adapted to 360° images without retraining," in *2023 8th Int. Conf. on Control and Robotics Engineering (ICCRE)*, 2023, pp. 78–83.

[33] H. Ishikawa, "Equilib: Equirectangular image library for 360° vision," 2023, accessed: 2025-08-04.

[34] S.-L. Hadj Sassia, T. Marsan, M. Benoussaad, and B. Watier, "A study on kinematic variabilities in human–human object co-manipulation," SSRN preprint, SSRN Electronic Journal, 2025, https://doi.org/10.2139/ssrn.5128137 (Accessed: 2025-08-04).

[35] C. Pizzolato, M. Reggiani, L. Modenese, and D. G. Lloyd, "Real-time inverse kinematics and inverse dynamics for lower limb applications using opensim," *Comput. Methods in Biomechanics and Biomedical Engineering*, vol. 20, no. 4, pp. 436–445, 2017, pMID: 27723992.

[36] Y. Xia, X. Zhou, E. Vouga, Q. Huang, and G. Pavlakos, "Reconstructing humans with a biomechanically accurate skeleton," in *Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, June 2025, pp. 5355–5365.

[37] S. Wang, J. Li, T. Li, Y. Yuan, H. Fuchs, S. De Mello, K. Nagano, and M. Stengel, "BLADE: Single-view Body Mesh Learning through Accurate Depth Estimation," in *arXiv*, 2024.