IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Calibration-Free Shoulder Kinematics Using Single and Dual Asynchronous RGB-D Cameras

VYTAUTAS ABROMAVIČIUS[1,2], JULIUS GRIŠKEVIČIUS[3], DALIUS MATUZEVIČIUS[2], (MEMBER, IEEE), ALGIRDAS MAKNICKAS[3], DARIUS PLONIS[2], (MEMBER, IEEE), AND RYTIS MASKELIŪNAS[1], (MEMBER, IEEE)

[1]Centre of Real Time Computer Systems, Faculty of Informatics, Kaunas University of Technology, 51423 Kaunas, Lithuania
[2]Department of Electronic Systems, Vilnius Gediminas Technical University, Plytinės g. 25, Vilnius, LT-10105, Lithuania
[3]Department of Biomechanical Engineering, Vilnius Gediminas Technical University, Plytinės g. 25, Vilnius, LT-10105, Lithuania

Corresponding author: Vytautas Abromavičius (e-mail: vytautas.abromavicius@ktu.lt).

This project has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-PD-24-29. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ABSTRACT This study experimentally evaluates whether shoulder joint kinematics can be accurately reconstructed using calibration free asynchronous RGB-D cameras under rehabilitation relevant conditions. We present a markerless framework that jointly learns temporal alignment, geometric consistency, and pose reconstruction using continuous-time modeling based on Neural Ordinary Differential Equations and implicit representations, eliminating the need for hardware synchronization or manual camera calibration. The system was validated in a controlled laboratory setting against a BTS Smart-DX optical motion capture reference during five clinically relevant shoulder movements. Performance was assessed for single- and dual-camera configurations. The dual-camera setup achieved a mean joint position error of 15.4 ± 2.8 mm with low temporal jitter (5.9 ± 0.7 mm), while the single-camera configuration showed reduced accuracy and higher sensitivity to occlusion. The results demonstrate that calibration-free asynchronous RGB-D systems can provide feasible shoulder kinematics, with a clear accuracy–complexity trade-off between single- and dual-camera deployments.

INDEX TERMS Computer vision, human pose estimation, motion capture, pose estimation, single person pose estimation.

NOMENCLATURE

A_{geo}	Geometric consistency matrix used in cross-modal attention.
β	Pose-dependent deformation parameters of the implicit human model.
C	Camera in a heterogeneous multi-camera system.
δ_i	Learned temporal offset (time-warping function) for camera i.
$\Delta_{\max}$	Maximum allowed temporal offset difference between camera streams.
$d_{2D}(\cdot)$	Distance between corresponding 2D observations across views.
$d_{3D}(\cdot)$	Distance between reconstructed 3D joint positions.
$f_\theta(\cdot)$	Neural Ordinary Differential Equation governing pose dynamics.
F_{ij}	Fundamental matrix relating cameras i and j.
$g_{\phi_i}(\cdot)$	Modality-specific measurement function for camera i.
$\gamma(\cdot)$	Fourier feature positional encoding function.
J	Number of skeletal joints.
K	Learned modality-specific feature weighting matrices.
L	Reconstruction loss.
M_i	Measurement acquired by camera i.
N	Number of cameras in the system.
P	3D human pose trajectory.
p	3D position of specific joint.
$q(\delta\|M)$	Variational posterior distribution over

	temporal offsets.
$R(\cdot)$	Regularization term.
Σ_i	Measurement noise covariance of sensor i.
Σ_ξ	Process noise covariance of pose dynamics.
τ	Continuous-time variable.
θ	Parameters of the pose dynamics network.
ϕ	Camera parameters of the measurement model.
ξ	Stochastic disturbance in the pose dynamics.

I. INTRODUCTION

In rehabilitation and clinical biomechanics, accurate motion capture is essential to assess patient progress, guide therapeutic interventions, and prevent injury [1]. Quantitative analysis of joint kinematics provides objective data that can assist in diagnosing musculoskeletal conditions, evaluating treatment outcomes, and refining rehabilitation programs [2]. Traditionally, such analyzes have been based on optical marker-based motion capture systems. These include multi camera setups that track reflective markers on the body and are widely regarded as the gold standard for measuring three dimensional movement [3]. Laboratory systems can achieve high accuracy, with only a few degrees of error in joint angle measurements [4], [5], allowing detailed biomechanical assessments for research and clinical decision making. However, marker-based systems also have well-documented limitations [6]. Attaching numerous markers to anatomical landmarks is time consuming and requires trained personnel. In addition, markers and associated equipment can hinder natural movement and introduce soft tissue artifacts that reduce the accuracy of the measurement [7]. Moreover, the cost and complexity of these systems confine their use to specialized motion laboratories, making them impractical for routine or field use. These limitations have driven the search for more accessible technologies to capture human movement in rehabilitation settings.

In recent years, markerless motion capture has emerged as a promising alternative that leverages video cameras or depth sensors, and computer vision algorithms to estimate human pose without physical markers [8]. Advances in depth sensing and real time pose estimation, including RGB-D cameras and deep learning models, have enabled markerless systems that track key body landmarks with reasonable accuracy in various environments [9]. Such systems are attractive for rehabilitation because they can allow for movement assessment in clinics or at home and provide immediate feedback to patients and clinicians. If sufficiently accurate, markerless setups could improve practicality and comfort by reducing preparation time and expanding access to high quality biomechanical data [10]. However, clinical adoption requires validation against established reference systems. Recent studies indicate that while markerless methods can capture basic spatiotemporal gait parameters, estimated joint angles may still deviate beyond clinically acceptable thresholds [11]. Therefore, a direct comparison with gold standard references remains necessary to establish accuracy and reliability [12]. Therefore, a rigorous evaluation under representative rehabilitation exercises is required to determine whether markerless solutions meet the precision requirements of clinical movement analysis.

The clinical deployment of markerless shoulder kinematics is limited less by the pose estimator capacity and more by the system level practicality [13]. In routine rehabilitation, solutions should operate without hardware synchronization and without manual camera calibration. They must also remain robust when fusing heterogeneous RGB and depth data in unconstrained environments. However, many multiview and implicit reconstruction pipelines assume synchronized and calibrated capture [14]. Heterogeneous sensor fusion methods often assume that temporal alignment is known or can be imposed externally.

To address these limitations, we introduce a framework that reformulates multimodal pose estimation using implicit neural representations. The method removes the need for hardware synchronization and explicit camera calibration by jointly learning temporal alignment, geometric consistency, and pose reconstruction in an end to end optimization. It is designed to be robust to heterogeneous camera configurations. Human motion is modeled as a continuous time process using Neural Ordinary Differential Equations, and the subject is represented as a deformable neural radiance field. This enables operation with flexible camera networks that incorporate available sensors, including stereo and infrared based depth. The system requires only two cameras with overlapping fields of view to recover full body pose.

II. HUMAN SKELETAL EVALUATION TECHNOLOGIES

The landscape of 3D human pose estimation has undergone transformative shifts with the advent of neural implicit representations and cross-modal learning, yet fundamental challenges in heterogeneous temporal synchronization remain largely unaddressed [15], [16]. The current literature indicates a clear gap. Methods developed for synchronized multiview systems are relatively mature, whereas approaches suited to practical unconstrained environments remain at an early stage of development.

Multiview pose estimation has evolved from traditional epipolar geometry and triangulation methods to sophisticated deep learning architectures. Recent transformer-based approaches have shown remarkable performance in multi-camera calibrated setups, some methods achieving millimeter-level accuracy under laboratory conditions [17], [18]. However, these advances

remain subject to important constraints. They typically assume precise hardware synchronization and known camera intrinsics and extrinsics. The recent trend toward temporal transformers and video pose estimation has improved temporal coherence, but exacerbates the synchronization dependency by requiring frame aligned sequences across views [19].

Neural Implicit Representations have emerged as a paradigm shifting approach for 3D reconstruction and novel view synthesis. The rapid evolution from NeRF to human specific implicit models has enabled unprecedented reconstruction quality [20], [21]. Recent works have integrated parametric body models with neural radiance fields, achieving photorealistic avatars from multiview inputs [22], [23]. However, these methods are dependent on synchronized captures and calibrated camera arrays. The latest dynamic NeRF variants handle temporal sequences but assume perfect frame synchronization, limiting their applicability to controlled studio environments [24].

Heterogeneous Sensor Fusion has seen an accelerated development with the proliferation of multimodal datasets. Cross-modal transformers and attention mechanisms have demonstrated impressive feature fusion capabilities, while sensor independent architectures have shown promise in maintaining performance under sensor stream errors or degradations [25], [26]. Recent work in infrared camera fusion and thermal RGB integration has pushed performance limits under challenging conditions [27]. Nevertheless, these approaches universally assume temporal alignment between modalities, either through hardware synchronization or manual postprocessing. The fundamental challenge of learning temporal relationships from unaligned heterogeneous streams remains unaddressed.

Continuous time representations have gained traction for handling irregularly sampled data. Neural ODEs and their recent extensions have shown promise in modeling continuous dynamics [28], while Fourier feature networks and positional encoding have enabled the representation of high-frequency signals [29]. Applications to human motion have demonstrated smooth interpolation capabilities, but these methods typically operate on prealigned sequences and lack explicit mechanisms for cross-sensor temporal alignment.

Self-supervised synchronization methods represent the most relevant adjacent research. Recent work has explored learning temporal offsets from video content, primarily for multiview action recognition [30]–[32]. However, these approaches focus on semantic alignment rather than geometric consistency and operate exclusively on RGB data. The extension to heterogeneous sensors and 3D geometric tasks remains unexplored.

Geometric Deep Learning advances have enabled for more robust multiview consistency. Differentiable triangulation methods and epipolar attention mechanisms have improved geometric reasoning, while learnable matching networks have improved the feature correspondence between views [33]. However, these geometric constraints are typically applied during post-synchronization rather than being integrated into the temporal alignment process itself.

The obvious research gaps emerge from this analysis. First, the temporal synchronization bottleneck remains. Most existing methods still depend on specialized hardware or manual alignment, which limits scalable deployment. Second, heterogeneous modality integration remains superficial, with most approaches performing late fusion rather than learning modality specific temporal and geometric relationships. Third, the theoretical foundations for identifiability in uncalibrated and unsynchronized systems are largely unexplored. Fourth, current evaluation paradigms focus on controlled settings, failing to assess real-world robustness to temporal misalignment and sensor heterogeneity.

Based on this analysis, a key unresolved question is whether a calibration-free and asynchronously captured RGB-D setup can achieve acceptable shoulder kinematics without relying on hardware synchronization or explicit camera calibration. In particular, it remains unclear how accuracy and temporal stability differ between minimal single-camera configurations and more robust dual-camera setups under rehabilitation-relevant movements. This study addresses this gap through an experimental validation against a gold-standard optical motion capture system, focusing on shoulder range-of-motion exercises and explicitly quantifying the trade-off between system complexity and kinematic accuracy. Moreover, beyond experimental validation, the work also contributes a probabilistic formulation for sensor fusion through the Heterogeneous Evidence Lower Bound (HELBO). In contrast to standard variational objectives, HELBO incorporates temporal uncertainty and geometric inconsistency directly into the model.

III. METHODS

A. STUDY DESIGN

Shoulder rehabilitation progresses through multiple stages, each characterized by different amplitudes of movement, pain levels, and compensatory strategies, making continuous monitoring challenging. The early stages involve limited, assisted, and pain-avoidant movements with high variability, while intermediate stages still exhibit compensations that obscure true shoulder kinematics. Fig 1 represents a controlled rehabilitation/assessment exercise in the later-stage, selected to ensure repeatable motion and reliable comparison with the gold-standard system. In this study, earlier stages were not addressed as the primary objective was methodological validation under conditions suitable for quantitative kinematic analysis.

FIGURE 1. Performing an exercise and recording the data in laboratory conditions

This study followed an experimental validation design to assess an advanced markerless pose estimation framework against a gold standard motion capture system. The evaluation was carried out in a controlled laboratory setting with two sequential phases. Phase 1 involved simultaneous recordings of human shoulder movements by the new markerless system and a reference BTS Bioengineering optical motion capture system, providing ground-truth joint kinematics for direct comparison. Phase 2 examined the performance of the markerless system in two configurations: a dual-camera (stereo) setup and a single camera setup to evaluate accuracy, robustness, and feasibility for rehabilitation use. By comparing markerless output from both configurations with reference data, we tested whether an asynchronous, calibration-free capture approach can achieve clinically acceptable accuracy for shoulder joint motion analysis. All trials were performed in the same lab environment with consistent procedures. Healthy adult volunteers (aged 20-40 years, without shoulder impairments) gave their informed consent and performed standardized movements. Participants wore normal athletic attire (with reflective markers only for the reference system) and no wearable sensors were required for the markerless method, minimizing encumbrance and preserving natural movement.

B. PARTICIPANTS

A cohort of healthy adult volunteers was recruited for laboratory trials. Participants were included if they had no known musculoskeletal or neurological impairments that affected shoulder motion. Each participant provided his informed consent in writing prior to participation, and the study protocol was approved by the institutional ethics committee. The characteristics of the participants are as follows: Volunteers (age range 20-40 years) were recreationally active and free of pain or limitations in movement in the upper extremities. To ensure safety and consistent performance, all participants received instructions and a brief training prior to data collection. Each volunteer wore form-fitting clothing to facilitate clear marker visibility for the reference system and unobstructed motion capture by the cameras. Informed consent was obtained from all subjects involved in the study. The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Vilnius Regional Biomedical Research Ethics committee (2023/6-1439-978).

C. MOVEMENT PROTOCOL

Participants performed a series of standardized shoulder movements that included the primary degrees of freedom of the shoulder joint. Five target motions were analyzed: shoulder flexion (extending the arm forward and overhead), shoulder extension (moving the arm backward), shoulder abduction (raising the arm laterally out to the side), shoulder internal rotation, and shoulder external rotation. Each movement was performed through the participant's comfortable full range of motion at a controlled speed, similar to clinical range of motion assessments. For dynamic movements such as flexion/extension and abduction, participants

raised and lowered their arm in a smooth, continuous motion. For rotational movements, the elbow was kept at 90° with the upper arm on the side, while the forearm rotated inward or outward. Each motion trial lasted approx. 10 to 15 s to capture a full cycle of movement. To improve reliability, three trials of each movement were recorded for each participant in each capture configuration. Participants were given short rest periods between tests to prevent fatigue. All movements were demonstrated and supervised by a researcher with a physical therapy background to ensure consistency with rehabilitative exercise techniques. A parcipant with markers performing an exercise is shown in Fig 1.

D. SYSTEM CONFIGURATION

The markerless motion capture system utilized two heterogeneous RGB-D cameras that operated without hardware synchronization or pre-calibration. We selected two consumer depth sensors representing different modalities: an Orbbec Femto Mega (indirect time-of-flight) and an Orbbec Gemini 2 (active stereo IR). Each device provides color and depth streams up to 30 Hz (with resolution 1280×800) and has a wide field of view (for example, 90° horizontal) covering a capture area 4×4 m. The cameras were mounted on tripods at approx. 1.2 m height (shoulder level), placed approximately 3 m from the subject at 45° to either side, resulting in overlapping views. Crucially, no multi-camera extrinsic calibration was performed; the relative poses of the cameras were not predetermined. Instead, the system treats each camera pose as an unknown that cannot be estimated during the learning process. Both RGB-D sensors were connected to a portable computing unit (Seeed reComputer J30 with NVIDIA Jetson Orin Nano, 8 GB), running Ubuntu 22.04 with JetPack 5.1. This edge GPU platform (approximately 40 TOPS AI performance) handled all data capture and on-device processing. The reference system was a BTS Smart-DX optical motion capture set-up with 8 infrared cameras (120 Hz). Retroreflective markers (9.5 mm) were placed on the bony landmarks of the shoulder and arm (modified Helen Hayes protocol). The BTS system, calibrated to submillimeter accuracy, provided ground-truth 3D trajectories of these markers for each trial. All sensors covered the same volume so that the subject remained within the view of both the depth cameras and the reference system throughout each movement.

E. DATA ACQUISITION AND SYNCHRONIZATION

Data from markerless cameras and the reference system were collected concurrently, but without direct electrical synchronization. Each Orbbec RGB-D camera ran on its own internal clock, streaming timestamped color and depth frames to the Nvidia Jetson for recording. A simple manual trigger (verbal 'start' cue) was used to begin recording on all devices at approximately the same time. Unlike the initial version of our setup, we did not employ a hardware trigger pulse or a unified clock signal to sync the cameras. Instead, any small temporal offsets between the streams were later corrected for by the continuous-time optimization framework itself. To facilitate an initial rough alignment, the participant performed a quick arm clap at the start of each trial (producing a distinctive motion and sound visible/audible to all systems); this served as a recognizable event for approximate synchronization, similar to a slate clap in filming. The neural ODE-based model inherently handles residual timing misalignment by adjusting the per-camera time offset parameters during training, effectively warping the timeline of each camera to maximize agreement. As an additional verification step after optimization, we cross-correlated the synchronized joint angle trajectories from the markerless and reference data, confirming that the recovered temporal alignment was accurate to within $5 - 10$ ms (comparable to the hardware sync tolerance in the original setup).

Spatial alignment between the markerless reconstruction and the reference coordinate system was performed in post-processing. Because the implicit framework estimates camera poses only up to an unknown global transform, we used a one time calibration motion at the start of each session to define a common reference frame. Participants assumed a T position with arms outstretched and performed predefined arm rotations while both systems recorded the motion. This sequence served as a dynamic calibration.

A best fit rigid transform, consisting of a 3D rotation and translation, was computed to map the markerless 3D joint locations to the BTS coordinate frame. The transform was estimated using the marker trajectories recorded during the calibration sequence of the T pose. The same transform was then applied to all subsequent markerless reconstructions. This enabled a direct comparison between the markerless joint positions and the reference measurements. The alignment did not require knowledge of the camera placement. It was computed solely from the data and was used only to express both outputs in a common coordinate system for evaluation.

The BTS system captured marker trajectories at 120 Hz. These data were downsampled to 30 Hz and time aligned to the markerless timeline using the learned temporal offset before error computation. All RGB-D sensor data and reference motion capture data were recorded for each trial and stored for offline analysis. No data loss or frame drops were observed during recording. The recorded data were then processed by the proposed pipeline to produce final 3D pose estimates, as described above. Finally, pose estimates and reference measurements were exported in C3D and CSV formats for quantitative analysis.

F. POSE ESTIMATION

We consider a challenging real-world scenario where N heterogeneous cameras $\mathcal{C} = \{C_1, C_2, \ldots, C_N\}$ are deployed in an unconstrained environment with unknown relative poses and unsynchronized temporal sampling. Each camera C_i $(i = 1, \ldots, N)$ acquires measurements $\mathbf{M}_i(t)$ at time t with an unknown temporal offset δ_i, representing the combined effects of clock drift, transmission delays, and variable processing latency. The 3D pose trajectory is $\mathbf{P}(t) = [\mathbf{p}_1(t)^\top; \ldots; \mathbf{p}_J(t)^\top] \in \mathbb{R}^{J\times 3}$, where $\mathbf{p}_j(t) \in \mathbb{R}^3$ is the Euclidean coordinate of joints j. The primary objective is to estimate the continuous 3D human pose trajectory $\mathbf{P}(\tau) \in \mathbb{R}^{J\times 3}$ from these asynchronous, multimodal observations without relying on external synchronization hardware or pre-calibration.

The fundamental observation model governing each camera's measurements is formalized as:

$$\mathbf{M}_i(t) = g_{\phi_i}(\mathbf{P}(t + \delta_i(t)), C_i) + \epsilon_i, \tag{1}$$

where g_{ϕ_i} denotes the modality-specific measurement function that maps the 3D pose to sensor observations, and $\epsilon_i \sim \mathcal{N}(0, \Sigma_i)$ represents additive measurement noise with modality-dependent covariance.

To address the inherent asynchrony across cameras, we model human pose evolution as a continuous-time dynamical system using Neural Ordinary Differential Equations (Neural ODEs). This formulation enables natural interpolation and extrapolation of poses at arbitrary timestamps, effectively decoupling pose estimation from discrete sampling rates. The pose dynamics are governed by:

$$\frac{d\mathbf{P}(\tau)}{d\tau} = f_\theta(\mathbf{P}(\tau), \tau) + \xi(\tau), \tag{2}$$

where f_θ is a neural network that learns the underlying dynamics of pose, and $\xi(\tau) \sim \mathcal{N}(0, \Sigma_\xi)$ captures the uncertainty in pose evolution, accounting for unpredictable human movement.

The joint optimization problem simultaneously estimates pose trajectories, temporal offsets, and model parameters:

$$\begin{aligned}\min_{\theta,\phi,\{\delta_i\}} \sum_{i=1}^{N} \mathcal{L}\left(\mathbf{M}_i(t), g_{\phi_i}(\mathbf{P}(t + \delta_i), C_i)\right) \\ +\lambda_1 R_{\text{pose}}(P, \theta) + \lambda_2 R_{\text{time}}(\{\delta_i\}),\end{aligned} \tag{3}$$

where $\phi = \{\phi_i\}$, the regularization terms R_{pose} and R_{time} enforce physiologically plausible poses and temporally smooth offset variations, respectively. Scalar weighting coefficients (hyperparameters) λ_1 and λ_2 balance the contributions of the regularization terms against the loss $\mathcal{L}$.

a: Modality-Specific Measurement Functions

Each sensor type employs a specialized measurement function that captures its unique physical characteristics and noise properties:

RGB Cameras leverage projective geometry combined with learned feature representations:

$$g_{\text{RGB}}(\mathbf{P}, C) = \Pi_C(\mathbf{P}) + \mathbf{K}_{\text{RGB}} \cdot \text{CNN}(\Pi_C(\mathbf{P})), \tag{4}$$

where Π_C denotes the camera projection operator and the CNN term captures appearance-based features. The matrix $\mathbf{K}$ represents the learned linear transformation weight that scales the feature representations.

Depth sensors utilize sparse point cloud measurements with signed distance fields:

$$\begin{aligned}g_{\text{depth}}(\mathbf{P}, C) =& \{\mathbf{p} \in \mathbf{P} : \rho(\mathbf{p}, S_{\text{depth}}) < \tau\} \\ &+ \mathbf{K}_{\text{depth}} \cdot \Phi_{\text{SDF}}(\mathbf{p}),\end{aligned} \tag{5}$$

where ρ measures point-to-scan distance and Φ_{SDF} represents the signed distance field, and S_{depth} maps joint skeleton to a surface (mesh).

Infrared Cameras model thermal radiation patterns and body temperature distributions:

$$\begin{aligned}g_{\text{IR}}(\mathbf{P}, C) =& T_{\text{body}}(\Pi_C(\mathbf{P})) \\ &+ \mathbf{K}_{\text{IR}} \cdot \Phi_{\text{thermal}}(\Pi_C(\mathbf{P})),\end{aligned} \tag{6}$$

where T_{body} captures body temperature variations and Φ_{thermal} models thermal radiation.

Stereo Cameras exploit binocular disparity constraints:

$$\begin{aligned}g_{\text{stereo}}(\mathbf{P}, C) =& \{\Delta(\Pi_{C_1}(\mathbf{P}), \Pi_{C_2}(\mathbf{P}))\} \\ &+ \mathbf{K}_{\text{stereo}} \cdot \Phi_{\text{disparity}}(\mathbf{P}),\end{aligned} \tag{7}$$

where Δ computes the disparity between stereo pairs and $\Phi_{\text{disparity}}$ encodes geometric constraints.

b: Geometric-Consistent Cross-Modal Attention

To effectively fuse information across heterogeneous sensors, we introduce a geometric-aware attention mechanism that explicitly incorporates multiview geometric constraints:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top + \mathbf{A}_{\text{geo}}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{8}$$

where d_k is a discrete time step. The geometric consistency matrix $\mathbf{A}_{\text{geo}}$ ensures that attention weights respect spatial relationships and prioritizes geometrically consistent feature correspondences across different viewpoints, effectively suppressing outliers and measurement noise:

$$\mathbf{A}_{\text{geo}}[i, j] = \exp\left(-\gamma \cdot \|\Pi_{C_i}(\hat{\mathbf{P}}) - \Pi_{C_j}(\hat{\mathbf{P}})\|^2\right), \tag{9}$$

where $\hat{\mathbf{P}}$ is the current human pose estimates of two devices i and j, while the parameter γ modulates the exponential decay rate.

c: Human-Centric Heterogeneous multimodal approach

We represent the human subject as a deformable neural radiance field that jointly models geometry, appearance, and pose-dependent deformations:

$$\sigma(\mathbf{x}), \mathbf{c}(\mathbf{x}) = \text{MLP}_\theta(\gamma(\mathbf{x}), t, \beta), \tag{10}$$

where $\sigma(\mathbf{x})$ represents volume density, $\mathbf{c}(\mathbf{x})$ denotes color, $\gamma(\mathbf{x})$ is positional encoding, and β encapsulates pose parameters. This implicit representation enables photorealistic rendering from arbitrary viewpoints while maintaining geometric consistency.

The multimodal rendering loss integrates information from all available sensors:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{IR}} + \mathcal{L}_{\text{stereo}}. \tag{11}$$

Each term is specifically designed to match the physical measurement principles of the corresponding sensor modality.

We employ Fourier Feature Flows to represent continuous-time pose trajectories, enabling smooth interpolation and natural motion priors:

$$\mathbf{P}(\tau) = \mathbf{P}(\tau_0) + \int_0^\tau f_\theta(\mathbf{P}(s), s)ds, \tag{12}$$

The Fourier feature encoding captures temporal patterns at multiple frequencies, and the spectral representation efficiently models both high-frequency motion details and low-frequency pose variations.

$$\begin{aligned}\gamma(\tau) = [&\sin(2\pi\omega_1\tau), \cos(2\pi\omega_1\tau), \ldots, \\ &\sin(2\pi\omega_k\tau), \cos(2\pi\omega_k\tau)].\end{aligned} \tag{13}$$

We derive a novel Heterogeneous Evidence Lower Bound (HELBO) that provides a principled probabilistic foundation for multimodal sensor fusion:

$$\begin{aligned}\log p(\mathbf{M}|\theta, \phi) \geq\ &\mathbb{E}_{q(\delta|\mathbf{M})}[\log p(\mathbf{M}|\mathbf{P}, \delta, \phi)] \\ &- D_{\text{KL}}(q(\delta|\mathbf{M})\|p(\delta)) - \lambda \cdot D_{\text{geo}}(\mathbf{P}),\end{aligned} \tag{14}$$

where $\delta = (\delta_1, \ldots, \delta_N)$.The geometric consistency term D_{geo} ensures the coherence of the multiview:

$$D_{\text{geo}}(\mathbf{P}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \|d_{3\text{D}}(\mathbf{P}) - d_{2\text{D}}(\mathbf{M}_i, \mathbf{M}_j)\|^2. \tag{15}$$

d: Pose Optimization

Camera-specific temporal warping functions automatically compensate for variable frame rates and transmission delays and once learned, these warping functions are used to enable non-linear time compensation, effectively synchronizing measurements without requiring external timing signals or manual calibration:

$$\tau_i(t) = t + \delta_i(t), \tag{16}$$

$$\delta_i(t) = \text{MLP}_{\delta_i}(t, \mathbf{M}_i(t)). \tag{17}$$

We formulate pose estimation as a constrained optimization problem solved via an Alternating Direction Method:

$$\begin{aligned}\min_{\mathbf{P}, \{\delta_i\}} \sum_{i=1}^{N} &\|\mathbf{M}_i(t) - g_{\phi_i}(\mathbf{P}(t + \delta_i(t)))\|^2 \\ &+ R_{\text{temporal}}(\mathbf{P}) + R_{\text{smooth}}(\delta_i),\end{aligned} \tag{18}$$

subject to physiologically plausible constraints to ensure temporal smoothness and prevent physically implausible pose transitions:

$$\|\delta_i - \delta_j\| < \Delta_{\max} \tag{19}$$

$$\|\mathbf{P}(t) - \mathbf{P}(t + \Delta t)\| < v_{\max} \cdot \Delta t. \tag{20}$$

Finally, the epipolar consistency loss enforces geometric coherence across all camera viewpoints:

$$\mathcal{L}_{\text{epipolar}} = \sum_{i \neq j}\sum_{j=1}^{N}\sum_{\mathbf{p} \in \mathbf{P}} \|\mathbf{p}_i^\top \mathbf{F}_{ij} \mathbf{p}_j\|^2, \tag{21}$$

where $\mathbf{F}_{ij}$ represents the fundamental matrix between cameras C_i and C_j, estimated directly from neural features. The approach ensures that the corresponding points in different perspectives satisfy the epipolar constraints, providing strong geometric regularization.

Under mild conditions requiring at least two cameras with overlapping fields of view and human motion exhibiting sufficient excitation, the parameters $\{\theta, \phi, \delta_i\}$ are identifiable up to the global scale and the temporal shift. It guaranties that the proposed method can uniquely recover both pose trajectories and temporal offsets from heterogeneous measurements. The approach relies on the observability of the system when human kinematic constraints are combined with multiview geometric constraints, ensuring that the optimization landscape contains a unique minimum corresponding to the ground truth solution.

G. EVALUATION METRICS

We evaluated the performance of the proposed system using a suite of standard metrics from the 3D human pose estimation literature, including accuracy, consistency, and efficiency.

The Mean per Joint Position Error (MPJPE) was used to measure the average Euclidean distance between each estimated joint position and the corresponding ground-truth position (from the BTS system) over all frames of a trial. It is reported in millimeters and reflects the overall 3D localization accuracy for body keypoints. We computed MPJPE for the major upper-body joints (shoulder, elbow, wrist) and averaged them to summarize the spatial error per trial.

Procrustes-Aligned MPJPE (PA-MPJPE) was used to isolate pose accuracy independent of global positioning, which is the MPJPE after a Procrustes alignment, meaning the predicted pose is optimally rigidly aligned

(rotation and translation, and optionally scale if necessary) to the ground-truth skeleton before error calculation. PA-MPJPE effectively removes any global offset or orientation mismatch, focusing on the accuracy of the relative joint configuration. This metric is useful for our calibration-free setup, since the markerless output may initially be in its own coordinate frame. A lower PA-MPJPE (compared to raw MPJPE) indicates that most of the errors are due to rigid misalignment rather than incorrect limb configuration.

The percentage of correct keypoints (PCK) and the area under the curve (AUC) were used to evaluate the fraction of keypoints that were estimated within a certain error threshold of the ground truth. PCK is defined as the percentage of joints with a position error below 50 mm. The AUC provides a single summary score that integrates PCK in a range of error tolerances (e.g. $0-100$ mm), offering a holistic view of accuracy. Higher AUC indicates better overall precision of keypoint localization.

Temporal Consistency (Jitter) was used to quantify the smoothness and stability of the estimated motion, and we measured the short-term temporal consistency of the joint trajectories. The metric was defined as the average frame-to-frame displacement of each joint after filtering out the participant's global motion (for example, subtracting the center-of-mass trajectory). Essentially, it captures the jitter or noise: Smaller values mean that the pose estimates from one frame to the next do not jump around and thus form a smooth trajectory. We report the mean per-frame jitter in millimeters for key joints as an indicator of output stability.

Finally, we recorded the end-to-end processing latency on the Jetson Orin Nano. The inference and update cycle operates at roughly $30-40$ ms per frame after convergence, which corresponds to an effective frame rate of about $25-30$ FPS. We report the average latency in milliseconds. The latency measurement covers all components of the system's computation on the edge device and thus indicates whether the approach could run live.

IV. RESULTS

Both markerless camera configurations successfully captured shoulder kinematics, with the dual-camera RGB-D system generally achieving higher accuracy and closer agreement with the reference standard. All values are reported as mean standard deviation ± between participants and trials, unless otherwise noted.

A. POSE ESTIMATION ACCURACY

Table 1 summarizes the performance of single-camera and dual-camera markerless configurations on various accuracy metrics and five different movement types (Flexion, Extension, Abduction, Internal rotation and External Rotation). The dual-camera system achieved significantly lower MPJPE compared to the single-camera setup (15.4 ± 2.8 mm vs. 26.5 ± 5.2 mm, $p < 0.001$). The Percentage of Correct Keypoints (PCK@50 mm) reached 98.9% for the dual-camera system, compared to 94.6% in the single-camera setting. The AUC also favored the dual setup (0.98 vs. 0.93). PA-MPJPE was also lower for the dual-camera system (13.7 mm vs. 20.8 mm overall).

The dual-camera RGB-D configuration leverages redundant viewpoints of the same motion. Multiview geometric consistency constraints reduce depth ambiguity and improve the stability of the reconstructed 3D skeleton. This reduces the misalignment of the global pose and improves the PCK/AUC. Continuous-time modeling with a Neural ODE estimates and corrects inter-camera temporal offsets (δ_i). The single-view setup does not provide these cross-view constraints and is more susceptible to depth noise, occlusion, and perspective-dependent scale errors, resulting in higher MPJPE.

B. TEMPORAL SMOOTHNESS

Table 2 summarizes joint position jitter for five types of movements. The dual-camera configuration produced consistently lower jitter than the single-camera setup in all motions. The average jitter decreased from 7.8 ± 1.1 mm (single camera) to 5.9 ± 0.7 mm (dual camera), with the greatest reductions observed during flexion and extension (from $8.2-8.4$ mm to $6.0-6.2$ mm).

The observed reduction indicates an improved temporal stability of the reconstructed joint trajectories. With two viewpoints, multiview geometric consistency suppresses depth-induced fluctuations and reduces the impact of transient occlusions. Continuous-time modeling further compensates for inter-camera temporal offsets δ_i, limiting frame-to-frame inconsistencies.

C. REAL-TIME PERFORMANCE

Both configurations were tested on an NVIDIA Jetson Orin Nano (8 GB) for latency and performance evaluation. Table 3 summarizes the processing latency per frame and the frame rate achieved. The single-camera system reached 29 Hz with an average latency of 34.5 ms. The dual-camera configuration, despite higher computational demands, maintained a frame rate at 22 Hz with a latency of 45.5 ms.

The additional computational cost is expected, as the dual-camera pipeline processes two RGB-D streams and performs cross-view operations (e.g., fusion and geometric consistency checks) in addition to temporal alignment. These steps improve robustness and accuracy, but introduce extra per-frame overhead relative to the single-view baseline.

Fig 2 depicts the angular trajectories (alpha, beta, gamma) of the shoulder joint estimated during five representative movements. The subplots correspond to different shoulder motions: (a) flexion, (b) extension,

TABLE 1. Accuracy metrics by movement type for both markerless configurations. Single vs dual camera (mean ± SD).

Movement	MPJPE (mm)		PA-MPJPE (mm)		PCK@50 mm (%)		AUC (0–1)	
	Single	Dual	Single	Dual	Single	Dual	Single	Dual
Flexion	28.8 ± 5.1	15.6 ± 2.4	21.8 ± 4.8	14.2 ± 2.3	93.2 ± 3.4	98.8 ± 1.1	0.923 ± 0.045	0.979 ± 0.012
Extension	28.4 ± 5.5	16.1 ± 2.6	22.2 ± 5.0	14.4 ± 2.4	92.5 ± 3.6	98.4 ± 1.0	0.920 ± 0.044	0.977 ± 0.013
Abduction	26.1 ± 4.7	14.9 ± 2.2	20.3 ± 4.6	13.3 ± 2.1	95.1 ± 3.0	99.1 ± 0.9	0.936 ± 0.038	0.985 ± 0.010
Internal Rotation	25.3 ± 4.5	15.1 ± 2.5	19.9 ± 4.4	13.5 ± 2.2	96.2 ± 2.7	99.0 ± 1.0	0.938 ± 0.036	0.982 ± 0.011
External Rotation	24.9 ± 4.2	14.7 ± 2.3	19.8 ± 4.3	13.2 ± 2.1	96.3 ± 2.9	99.3 ± 0.8	0.941 ± 0.035	0.986 ± 0.009
Overall	**26.5 ± 5.2**	**15.4 ± 2.8**	**20.8 ± 5.0**	**13.7 ± 2.5**	**94.6 ± 3.1**	**98.9 ± 1.0**	**0.930 ± 0.040**	**0.980 ± 0.010**

TABLE 2. Joint position jitter (mm) by movement type.

Movement	Single Camera (mm)	Dual Camera (mm)
Flexion	8.2 ± 1.0	6.0 ± 0.6
Extension	8.4 ± 1.2	6.2 ± 0.7
Abduction	7.9 ± 1.1	5.8 ± 0.6
Internal Rotation	7.5 ± 1.0	5.7 ± 0.5
External Rotation	7.3 ± 0.9	5.6 ± 0.5
Average	**7.8 ± 1.1**	**5.9 ± 0.7**

TABLE 3. Computational performance comparison between single- and dual-camera configurations.

Metric	Single Camera	Dual Camera
Avg. Latency (ms/frame)	34.5 ± 2.3	45.5 ± 3.1
Frame Rate (Hz)	29.0	22.0

(c) abduction, (d) internal rotation, and (e) external rotation. Each line style represents one of the Euler angles that describe the humeral orientation. The plots reveal consistent periodic patterns for large-range movements (a-c), while rotational movements (d, e) exhibit smaller amplitudes and stronger coupling between Beta and Gamma. Overall, the results demonstrate that the system robustly reconstructs shoulder kinematics across various types of movement, with reduced reliability primarily in axial rotation and occlusion-prone configurations.

Practically, small fluctuations in depth or partial occlusions of distal landmarks can induce compensatory changes across multiple Euler channels, even when the underlying physical rotation is predominantly axial. The dual-view constraint reduces, but cannot eliminate this effect, because axial rotation can remain weakly observable from external cameras when the arm is close to the trunk.

Fig 3 shows a comparison of the shoulder flexion angle trajectories estimated by RGB-D systems with single and dual cameras against the reference based on the BTS marker (dashed). The subplots include single (a) and dual (b) samples to illustrate the strengths and limitations of the proposed approach. The well-aligned cases demonstrate that the dual-camera configuration produces smooth and temporally consistent trajectories closely matching the reference, even during dynamic motion phases. Moreover, the learned per-camera time offsets δ_i are better constrained because cross-view agreement provides a strong temporal cue. In contrast, the single camera exhibits increased noise, local misalignment, and occasional phase lag, highlighting the influence of depth noise.

D. ABLATION STUDY

To assess the contribution of specific architectural elements, we conducted an ablation study comparing full model performance with versions with selectively disabled features. Table 4 presents the impact on MPJPE and jitter when disabling (1) neural ODEs (replaced by discrete interpolation), (2) geometric attention and (3) time-warping compensation.

Removing neural ODE-based continuous-time modeling increased MPJPE by more than 20%, confirming its importance for robust motion interpolation. Disabling geometric attention led to a 10% increase in jitter and angle error, likely due to reduced view consistency. Removing time warping (that is, relying on raw timestamps) caused a significant degradation in accuracy and smoothness, underscoring the model's reliance on learned alignment for asynchronous fusion.

V. DISCUSSION

Recent advances in markerless pose estimation offer several points of comparison to our calibration-free RGB-D implicit neural framework. Traditional 2D pose estimators like OpenPose and MoveNet require only a single RGB camera and no explicit calibration, making them popular baselines for in the wild use. For example, OpenPose and Google MoveNet (Thunder) have demonstrated the ability to measure gait kinematics with mean errors of the order of only 4 to 5° at the hip and knee angles [34]. In fact, OpenPose was reported to achieve a 5.1° error for knee flexion during gait, outperforming other open source models in that scenario [35]. These 2D frameworks are lightweight and real-time, but they output either 2D joint locations or require heuristic depth estimation to infer 3D. In rehabilitation settings where absolute 3D accuracy is needed, multi-camera systems are often employed.

Our implicit approach, which does not require prior camera calibration or hardware sync, further reduces

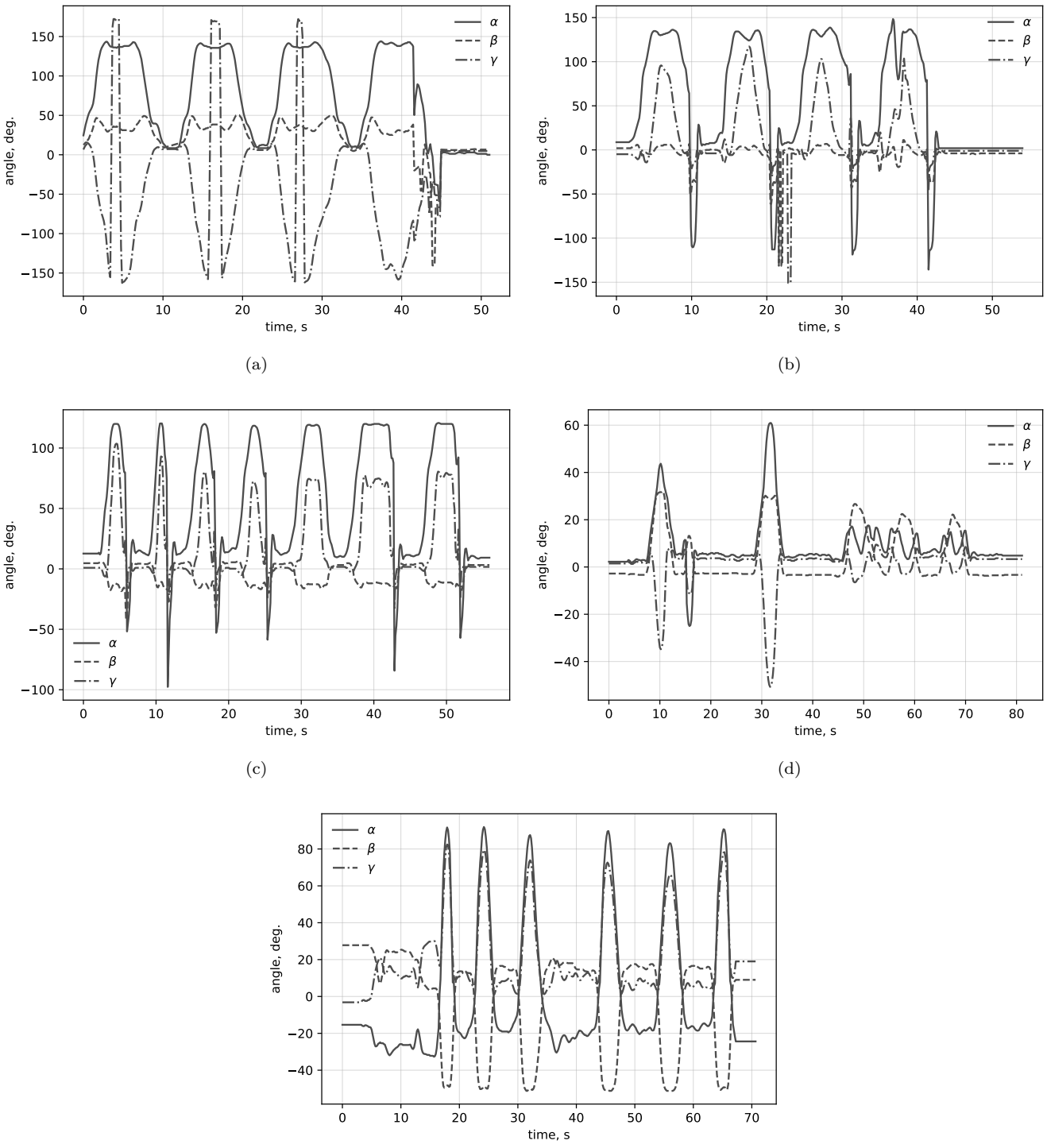
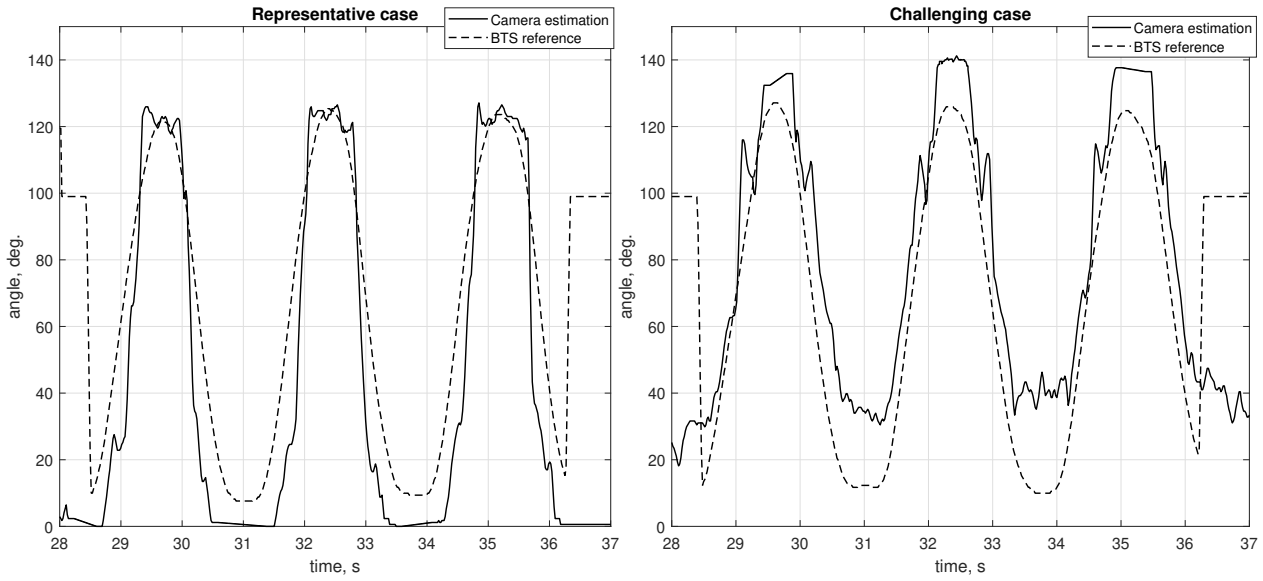
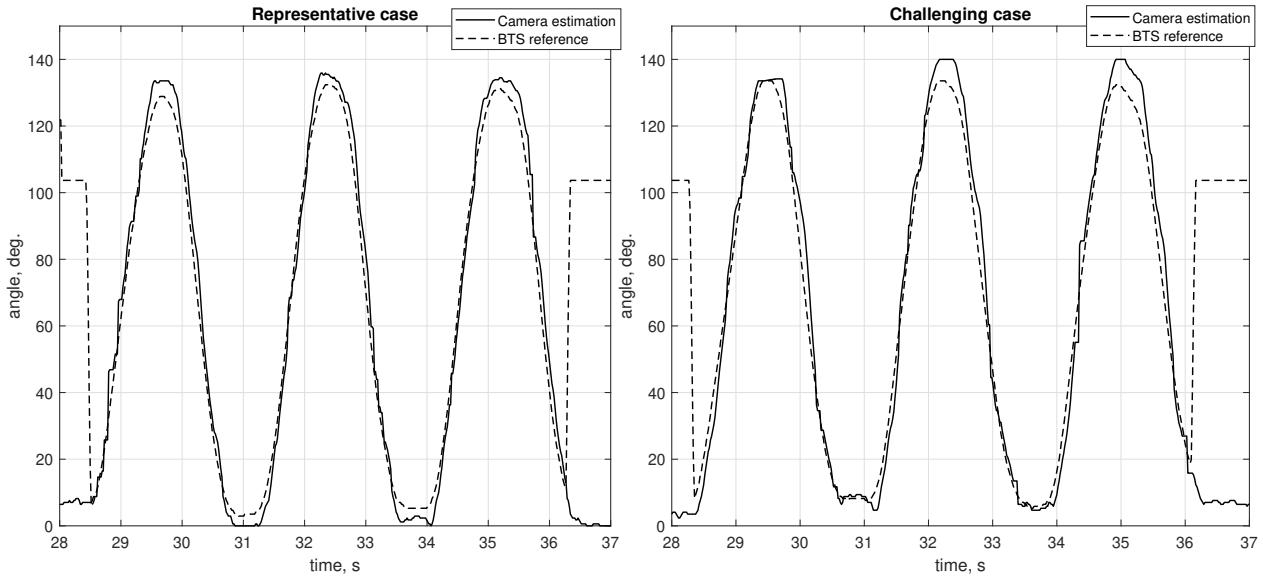


FIGURE 2. Angular trajectories (alpha, beta, gamma) of the shoulder joint estimated during five representative shoulder movements: flexion (a); extension (b); abduction (c); internal rotation (d); external rotation (e).

the error. In our laboratory tests, the dual RGB-D camera configuration reached MPJPE 15.4 mm (vs. 26.5 mm with a single camera). This is a substantial improvement, indicating that our NeRF+Neural ODE method can leverage the multiview geometry without explicit calibration, outperforming even some calibrated multi-camera methods in accuracy. This performance is competitive with established multiview pipelines that



(a)



(b)

FIGURE 3. Comparison of the shoulder flexion angle trajectories estimated by RGB-D systems (solid line) with single (a) and dual cameras (b) against the reference based on the BTS marker (dashed).

TABLE 4. Ablation results of proposed model vs. each feature disabled.

Configuration	MPJPE (mm)	PCK @ 50 mm (%)	Jitter (mm)
Full Model (Ours)	15.4	98.9	5.9
w/o Neural ODE	18.2	96.5	7.1
w/o Geometric Attention	16.7	97.2	7.6
w/o Time Warping Compensation	21.4	94.3	8.2

typically assume calibrated, synchronized cameras. For example, RANSAC-based OpenPose triangulation has been reported in 22.8 mm MPJPE in Human3.6M [36].

Compared with temporal smoothness, Buker et al [37] quantified repeatability-type variability of Azure Kinect body tracking under controlled conditions using a mannequin and reported median per-recording errors of $1.41 - 1.47$ mm across a set of key joints, but sub-

stantially larger errors for distal joints, for foot markers $4.56-6.11$ mm across repeated recordings. In particular, our dual-camera jitter ($5.6-6.2$ mm) is comparable under static conditions. However, our single camera jitter ($7.3-8.4$ mm) exceeds it, suggesting instability introduced by single view depthor transient occlusions during dynamic shoulder range of motion. At sensor level, Kurillo et al. [38] reported random depth error (standard deviation across 1000 frames) of $0.6-3.7$ mm, indicating that millimeter-scale depth noise can readily propagate into multi-millimeter 3D joint jitter once kinematic inference and occlusion handling are involved.

The measured real-time performance is relatively competitive with similar pose estimation pipelines, where optimized lightweight models reached approx. 50 Hz [39] or 43 Hz [40] frame rates. However, these results typically exclude the additional cost of RGB-D fusion, cross-view geometric consistency, and temporal alignment that are intrinsic to our proposed dual-stream 3D setting.

Empirically, many RGB-D and markerless upper-limb validation studies report that internal/external rotation is the least accurate shoulder angle. For example, Cai et al. reported good waveform agreement for flexion/extension, moderate for adduction, and poor for internal/external rotation of the shoulder using Kinect V2 against marker-based motion capture [41]. In addition, recent multicamera pose fusion system targets robustness under partial visibility [42], reinforcing cross-view consistency, while suppressing trajectory noise.

Unlike 2D pose estimators, our method produces true 3D trajectories and continuous temporal alignment. We also note that emerging NeRF-based pose systems are conceptually similar to our approach. For example, A-NeRF employs an articulated NeRF to jointly learn a deformable body model and refine pose estimates from single or multiple views without tedious camera calibration [43]. Such systems demonstrate that implicit volumetric representations can align multiview data in an ad hoc setup, much like our framework. Furthermore, a very recent study used human motion itself as a calibration cue for unsynchronized, uncalibrated videos, optimizing camera poses and time offsets in a dynamic NeRF pipeline [44]. Top state-of-the-art approaches highlight a trend: using learned models of human shape and motion (NeRFs, skeleton priors, etc.) can compensate for the lack of calibration, turning human subjects into their own calibration targets. Our work fits squarely in this trend, combining an implicit Neural Radiance Field for the subject with a Neural ODE temporal model, and compares favorably with the accuracy levels of the best calibration-free systems currently available.

A. LIMITATION AND FUTURE WORK

A limitation of this study is that only controlled shoulder movements representative of later-stage rehabilitation or functional assessment were evaluated. Earlier rehabilitation stages often involve restricted, assisted, or pain-limited motions with pronounced compensatory strategies, which pose additional challenges for reliable kinematic monitoring, especially for markerless systems. Consequently, the presented results cannot be directly generalized to early-stage clinical rehabilitation. Future work will extend validation to earlier rehabilitation phases and patient populations, focusing on robustness to limited range of motion, atypical movement patterns, and pain-related adaptations.

In addition, the evaluation involved only healthy adults performing shoulder movements, providing a controlled setting but limiting generalizability to clinical populations. Patients with abnormal synergies, assistive devices, or pain-related adaptations may deviate from the motion priors learned by our model. Future work should include patient data, additional joints, and varied exercises. The current model training is session-specific and requires optimization per participant, which is time-consuming and reduces generalizability. Adopting quick-or-fast-adaptation methods could enable faster initialization for new subjects. The system also assumes a single subject and static background, posing challenges for multiperson or dynamic environments typical in home settings. Moreover, while the dual-camera setup offers high accuracy, the single-camera mode shows higher error and occlusion sensitivity, warranting improvements through shape priors or inertial data. Lastly, the implicit model lacks interpretability compared to explicit kinematic models, which may hinder clinical acceptance. Hybrid implicit-explicit approaches could enhance both accuracy and explainability.

The proposed framework regulates the influence of heterogeneous sensor streams implicitly through probabilistic modeling, geometric consistency, and attention-based fusion, rather than by enforcing explicit contribution constraints. This design allows the optimization to adaptively emphasize reliable observations while suppressing inconsistent ones. More explicit mechanisms for contribution balancing, such as bounded weighting or constraint-based regularization, could further improve numerical conditioning and stability in extreme cases of noise imbalance or sensor degradation. However, it may also reduce flexibility by limiting the model's ability to disregard severely corrupted measurements. The observed stability and accuracy indicate that implicit regulation is sufficient in the evaluated setting, while more constrained formulations remain an interesting direction for future extensions.

Furthermore, future research could explore the integration of additional sensing modalities to enhance robustness and accuracy. Our framework could naturally incorporate thermal cameras or IMUs: thermal data could aid in visual occlusion tracking, while inertial signals could guide pose dynamics within the Neural

ODE. Another direction involves reducing dependence on labeled data through self- or semi-supervised learning, active learning, or synthetic data generation, enabling adaptation to new environments with minimal annotation. Computational efficiency also remains a priority: although the trained model frame rate was 22 Hz, higher resolutions or complex scenes may require model compression or faster ODE solvers to enable deployment on consumer devices. Overall, this study demonstrates a proof-of-concept for calibration-free markerless motion capture. Enhancing generalization, efficiency, and real-world validation are key steps toward translating it into a practical rehabilitation tool.

VI. CONCLUSION

This work experimentally validated a calibration-free markerless framework for shoulder kinematics estimation using asynchronous RGB-D cameras. Without relying on hardware synchronization or explicit camera calibration, the proposed approach achieved subcentimeter-level joint position accuracy and low temporal jitter when using a dual-camera configuration, as verified against a BTS Smart-DX reference system. Although a single-camera setup retained practical usability, it exhibited higher errors and reduced robustness in movements involving occlusion or axial rotation. Ablation results demonstrated that continuous-time modeling, learned temporal alignment, and multiview geometric constraints are essential to achieve stable and accurate reconstruction. These findings indicate that high-quality shoulder kinematic assessment can be performed outside specialized motion laboratories with minimal setup, provided that at least two unsynchronized RGB-D cameras are available. Future work should extend validation to clinical populations, earlier rehabilitation stages, and faster subject adaptation to further support real-world deployment.

The findings indicate that high-fidelity analysis of the above shoulder movements can be performed outside laboratory settings with minimal setup, allowing practical rehabilitation monitoring and biofeedback. The main limitations are evaluation on healthy adults and session-specific optimization under single-subject assumptions. Future work should extend validation to clinical cohorts, broaden the movement repertoire beyond the five tested actions, possibly handle multiperson scenes, and investigate faster adaptation and additional sensing (e.g., inertial or thermal) within the same implicit and continuous framework.

REFERENCES

[1] A. Andres, M. Roland, M. Orth, and S. Diebels, "From injury to full recovery: Monitoring patient progress through advanced sensor and motion capture technology," *Sensors*, vol. 25, no. 13, p. 3853, 2025.

[2] A. de los Reyes-Guzmán, I. Dimbwadyo-Terrer, F. Trincado-Alonso, F. Monasterio-Huelin, D. Torricelli, and A. Gil-Agudo, "Quantitative assessment based on kinematic measures of functional impairments during upper extremity movements: A review," *Clinical Biomechanics*, vol. 29, no. 7, pp. 719–727, 2014.

[3] H. Noorbhai, S. Moon, and T. Fukushima, "A conceptual framework and review of multi-method approaches for 3d markerless motion capture in sports and exercise," *Journal of sports sciences*, vol. 43, no. 12, pp. 1167–1174, 2025.

[4] A. Alahmari, L. Herrington, and R. Jones, "Concurrent validity of two-dimensional video analysis of lower-extremity frontal plane of movement during multidirectional single-leg landing," *Physical Therapy in Sport*, vol. 42, pp. 40–45, 2020.

[5] R. M. Maura, S. Rueda Parra, R. E. Stevens, D. L. Weeks, E. T. Wolbrecht, and J. C. Perry, "Literature review of stroke assessment for upper-extremity physical function via eeg, emg, kinematic, and kinetic measurements and their reliability," *Journal of NeuroEngineering and Rehabilitation*, vol. 20, no. 1, p. 21, 2023.

[6] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "Metrabs: metric-scale truncation-robust heatmaps for absolute 3d human pose estimation," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 16–30, 2020.

[7] Z. Niu, K. Lu, J. Xue, X. Qin, J. Wang, and L. Shao, "From methods to applications: A review of deep 3d human motion capture," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 11 340–11 359, 2024.

[8] W. W. Lam, Y. M. Tang, and K. N. Fong, "A systematic review of the applications of markerless motion capture (mmc) technology for clinical measurement in rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 20, no. 1, p. 57, 2023.

[9] R. Bashirov, A. Ianina, K. Iskakov, Y. Kononenko, V. Strizhkova, V. Lempitsky, and A. Vakhitov, "Real-time rgbd-based extended body pose estimation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2807–2816.

[10] C. Halmich, L. Höschler, C. Schranz, and C. Borgelt, "Data augmentation of time-series data in human movement biomechanics: A scoping review," *PloS one*, vol. 20, no. 7, p. e0327038, 2025.

[11] V. Abromavičius, E. Gisleris, K. Daunoravičienė, J. Žižienė, A. Serackis, and R. Maskeliūnas, "Enhanced human skeleton tracking for improved joint position and depth accuracy in rehabilitation exercises," *Applied Sciences*, vol. 15, no. 2, p. 906, 2025.

[12] R. Maskeliūnas, A. Kulikajevas, R. Damaševičius, J. Griškevičius, and A. Adomavičienė, "Biomac3d: 2d-to-3d human pose analysis model for tele-rehabilitation based on pareto optimized deep-learning architecture," *Applied sciences*, vol. 13, no. 2, p. 1116, 2023.

[13] W. van den Hoorn, A. Fabre, G. Nardese, E. Y.-S. Su, K. Cutbush, A. Gupta, and G. Kerr, "The future of clinical active shoulder range of motion assessment, best practice, and its challenges: narrative review," *Sensors*, vol. 25, no. 3, p. 667, 2025.

[14] N. Wilser, F. Cordier, and Y. Maillot, "A survey of 3d human body reconstruction from single and multiple camera views," *International Journal of Image and Graphics*, p. 2650014, 2024.

[15] A. F. R. Nogueira, H. P. Oliveira, and L. F. Teixeira, "Markerless multi-view 3d human pose estimation: A survey," *Image and Vision Computing*, p. 105437, 2025.

[16] Y. Liu, C. Qiu, and Z. Zhang, "Deep learning for 3d human pose estimation and mesh recovery: A survey," *Neurocomputing*, vol. 596, p. 128049, 2024.

[17] H. Ma, L. Chen, D. Kong, Z. Wang, X. Liu, H. Tang, X. Yan, Y. Xie, S.-Y. Lin, and X. Xie, "Transfusion: Cross-view fusion with transformer for 3d human pose estimation," *arXiv preprint arXiv:2110.09554*, 2021.

[18] H. Shuai, L. Wu, and Q. Liu, "Adaptive multi-view and temporal fusing transformer for 3d human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4122–4135, 2022.

[19] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12 922–12 943, 2023.

[20] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in neural information processing systems*, vol. 33, pp. 7462–7473, 2020.

[21] B. Jiang, X. Ren, M. Dou, X. Xue, Y. Fu, and Y. Zhang, "Lord: Local 4d implicit representation for high-fidelity dynamic human modeling," in *European Conference on Computer Vision*. Springer, 2022, pp. 307–326.

[22] X. Zhou, S. Peng, Z. Xu, J. Dong, Q. Wang, S. Zhang, Q. Shuai, and H. Bao, "Animatable implicit neural representations for creating realistic avatars from videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4147–4159, 2024.

[23] Y. Huang, H. Yi, W. Liu, H. Wang, B. Wu, W. Wang, B. Lin, D. Zhang, and D. Cai, "One-shot implicit animatable avatars with model-based priors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8974–8985.

[24] Y. Chen, Y. Zhan, Z. Zhong, W. Wang, X. Sun, Y. Qiao, and Y. Zheng, "Within the dynamic context: Inertia-aware 3d human modeling with pose sequence," in *European Conference on Computer Vision*. Springer, 2024, pp. 491–508.

[25] W.-Y. Lee, L. Jovanov, and W. Philips, "Cross-modality attention and multimodal fusion transformer for pedestrian detection," in *European conference on computer vision*. Springer, 2022, pp. 608–623.

[26] K. Li, J. Li, D. Guo, X. Yang, and M. Wang, "Transformer-based visual grounding with cross-modality interaction," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 6, pp. 1–19, 2023.

[27] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *IEEE Transactions on intelligent transportation systems*, vol. 24, no. 12, pp. 14 679–14 694, 2023.

[28] C. Zang and F. Wang, "Neural dynamics on complex networks," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 892–902.

[29] Y. Li, S. Si, G. Li, C.-J. Hsieh, and S. Bengio, "Learnable fourier features for multi-dimensional spatial positional encoding," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 816–15 829, 2021.

[30] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[31] L. Yin, R. Han, W. Feng, and S. Wang, "Self-supervised human pose based multi-camera video synchronization," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1739–1748.

[32] V. Abromavičius, E. Gisleris, K. Daunoravičienė, J. Žižienė, A. Serackis, and R. Maskeliūnas, "Robust skeletal motion tracking using temporal and spatial synchronization of two video streams," *Plos one*, vol. 20, no. 8, p. e0328969, 2025.

[33] N. Heidari and A. Iosifidis, "Geometric deep learning for computer-aided design: A survey," *IEEE Access*, 2025.

[34] E. P. Washabaugh, T. A. Shanmugam, R. Ranganathan, and C. Krishnan, "Comparing the accuracy of open-source pose estimation methods for measuring gait kinematics," *Gait & posture*, vol. 97, pp. 188–195, 2022.

[35] F. Roggio, B. Trovato, M. Sortino, and G. Musumeci, "A comprehensive analysis of the machine learning pose estimation models used in human movement and posture analyses: A narrative review," *Heliyon*, vol. 10, no. 21, 2024.

[36] K. Bae, S. Lee, S.-Y. Bak, H. S. Kim, Y. Ha, and J. H. You, "Concurrent validity and test reliability of the deep learning markerless motion capture system during the overhead squat," *Scientific Reports*, vol. 14, no. 1, p. 29462, 2024.

[37] L. Büker, M. Hackbarth, V. Quinten, A. Hein, and S. Hellmers, "Towards comparable quality-assured azure kinect body tracking results in a study setting—influence of light," *Plos one*, vol. 19, no. 8, p. e0308416, 2024.

[38] G. Kurillo, E. Hemingway, M.-L. Cheng, and L. Cheng, "Evaluating the accuracy of the azure kinect and kinect v2," *Sensors*, vol. 22, no. 7, p. 2469, 2022.

[39] L. Liu, E. B. Blancaflor, and M. Abisado, "A lightweight multi-person pose estimation scheme based on jetson nano," *Applied Computer Science*, vol. 19, no. 1, pp. 1–14, 2023.

[40] C. Neff, A. Sheth, S. Furgurson, J. Middleton, and H. Tabkhi, "Efficienthrnet: efficient and scalable high-resolution networks for real-time multi-person 2d human pose estimation," *Journal of Real-Time Image Processing*, vol. 18, no. 4, pp. 1037–1049, 2021.

[41] L. Cai, Y. Ma, S. Xiong, and Y. Zhang, "Validity and reliability of upper limb functional assessment using the microsoft kinect v2 sensor," *Applied bionics and biomechanics*, vol. 2019, no. 1, p. 7175240, 2019.

[42] L. Bragagnolo, M. Terreran, D. Allegro, and S. Ghidoni, "Multi-view pose fusion for occlusion-aware 3d human pose estimation," in *European Conference on Computer Vision*. Springer, 2024, pp. 117–133.

[43] S.-Y. Su, F. Yu, M. Zollhöfer, and H. Rhodin, "A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose," *Advances in neural information processing systems*, vol. 34, pp. 12 278–12 291, 2021.

[44] S. Kim, J. Bae, Y. Yun, H. Lee, G. Bang, and Y. Uh, "Sync-nerf: Generalizing dynamic nerfs to unsynchronized videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2777–2785.