

From Pixels to Performance: A Survey of the Estimation-to-Biomechanics Pipeline for Vision-Based Human Pose Estimation in Sports and Movement Health

Sushma Zade, Vedika Andhale, Aman Mishra, Vishal Pandit, and Pratik Meshram

Department of Information Technology

Sandip Institute of Technology and Research Centre (SITRC), Nashik, Maharashtra, India

sushma.zade@sitrc.org; vedikaandhale.112023344@sitrc.org; amanmishra.112023022@sitrc.org;

vishalpandit.112023122@sitrc.org; meshrampratik.112023329@sitrc.org

Abstract: Traditional quantitative movement analysis relies either on the subjective visual observations of a coach or laboratory-based motion capture involving markers. Human Pose Estimation (HPE) using deep learning could provide a third solution where markers can be omitted, and anatomical landmarks can be extracted from regular RGB videos through cameras, leading to movement analysis in non-laboratory environments. The potential offered by HPE is sometimes exaggerated relative to empirical results. This survey reviews 45 articles, published from 2021 to 2026, covering the four stages of the process from video to movement analysis: landmark estimation, transformation of coordinates to biomechanical metrics, sport-specific analysis, and validation against a gold standard. Our contributions include a taxonomy based on the pipeline, a comparative study of landmark estimation and metric transformation approaches, landmark-to-joint angle error propagation analysis, and temporal validity budget modeling. Three results come through clearly. First, accuracy is the best-optimized and most widely reported metric, but it is not the constraint driving adoption. Second, the translation and validation phases combine to occupy less than one fifth of the corpus, despite being essential to every practical application. Third, the reported accuracy of different frameworks is not comparable between implementations of the same framework. We summarize six converging gaps, and present our reference implementation on consumer hardware, including performance metrics (CPU throughput of 29.2 frames per second and GPU throughput of 74.1 frames per second for a 33-landmark estimator; CPU throughput of 4.4 frames per second and GPU throughput of 102.0 frames per second for a 19-landmark estimator), and what these metrics demonstrate and what they do not demonstrate

Keywords: Human pose estimation, markerless motion capture, sports biomechanics, movement health, joint kinematics, validation, edge inference, systematic review

I. INTRODUCTION

Sports and clinical movement analysis both have the same challenge with regards to measuring certain parameters. Joint angles, angular velocities, timing of phase and left and right sidedness are parameters of interest in biomechanical science; however, the acquisition of these parameters was dependent on visual analysis by the experts or by instrumented data collection. Visual observation of experts is cheap and scaleable but not quantifiable and subject to errors because of bias. Marker based optical motion capture systems are accurate but need instrumented athletes and



multi-camera volume calibration in indoor conditions. The result is that quantification of movements stays in the domain of elite sports and gait clinics only.

Human pose estimation using deep learning provides a third approach. Human pose estimation estimates the location of anatomical landmarks like shoulders, elbows, wrists, hips, knees, and ankles. Since it uses standard RGB images, it enables an inexpensive, non-marker-based, camera-based pipeline where biomechanically meaningful parameters can be extracted from video footage captured by a smartphone. The same technology can find applications in health informatics, where tasks such as remote gait evaluation after orthopaedic surgery, monitoring adherence to telerehabilitation, screening for certain gait patterns linked with anterior cruciate ligament injury, and long-term monitoring of gait deterioration due to neurological conditions would need the same features.

The interest has increased at a fast rate. There have been four systematic reviews of pose estimation for sports and physical activities, independently carried out between 2021 and 2025 [1]–[4], all with a unique perspective on the topic: computer vision, sports science, data collection, and commercial applications. At the same time, general pose estimation has made progress at a rate faster than the demand of the applied research [5]–[7].

A. Problem and Motivation

One thing that repeatedly emerges from reviewing literature recently published on this topic is the fact that existing applications of pose estimation algorithms in sports are mostly designed as single-exercise pipelines. An isolated system is created for a single exercise, which involves similar geometric logic, which has to be reinvented for the second exercise [1], [2]. Uçar et al. [3] found that the datasets that are currently used tend to be biased towards team sports rather than individual sports. Edriss et al. [4] reveal that the accuracy of pose estimation algorithms largely depends on how a system is implemented, where cameras are located, their frame rate, and lighting conditions.

Concurrently, the literature on architectural design has largely focused only on maximizing the geometric fidelity of predictions on laboratory datasets. Transformer-based 3D pose-lifting models [8], [9] set new state-of-the-art performance on Human3.6M and MPI-INF-3DHP, but they depend upon a fixed-length window of past and future frames, creating structural latency inconsistent with feedback in real time, and they require inference on the GPU. None of them establishes a convention for interpreting the recovered coordinates as biomechanical quantities.

In this case, the problem does not lie in a lack of estimators. The problem is the absence of a disciplined intermediate layer: a standardized way to convert landmark coordinates into biomechanical quantities, the conditions under which such quantities can be said to be accurate, and an evaluation against an independent reference standard. This review centers around that intermediate layer and its neighbours.

B. Health-Informatics Relevance

Considering the problem only in terms of sports performance overlooks its broader implications. In the domain of movement health, the same pipeline is used for four distinct purposes with varying error tolerance. *Screening* involves checking if a movement pattern lies outside of a normative range and allows for considerable error tolerance if the range is broad. *Monitoring* involves comparing a quantity with a previously recorded value from the same individual, where excellent test-retest reproducibility is required, but not necessarily accuracy. *Assessment* involves checking consistency with a laboratory standard, where both are necessary. *Feedback* involves providing feedback during the performance of a movement, where the latency requirement is the most stringent. Confusion between the four tasks is a frequent cause of exaggerated claims, where a system designed for monitoring is claimed to have been validated for assessment.

There are two more issues particular to the use of AI in healthcare. The first one is the fact that the video of an individual is personal information, and medical video involves higher sensitivity, which makes on-device inference a completely different thing from uploading video files to a server. The second one is that the measurement may be accurate on average but inaccurate for a particular subset of people based on their body shape, skin color, clothes, or any assistive devices they might have.



C. Contributions

This survey merges two threads of research—the advancement of architectural approaches and movement analysis techniques—with the intermediary and ultimate steps bridging them. What we contribute is:

- 1) **A pipeline-based taxonomy.** Instead of categorizing the literature on the basis of sports, algorithm families, or sensing devices, we arrange the 45 reviewed papers into the four-stage pipeline that transforms video footage into coaching and clinical advice through: (I) landmark detection, (II) transformation into biomechanical measurements, (III) application in a particular sport, and (IV) validation and implementation. This dimension is more helpful for a practitioner than sports or architecture, since it reveals at which stage each thread contributes and where the gaps are.
- 2) **Comparative synthesis of the two important design choices.** Comparisons of landmark-estimation backbone systems for accuracy, latency, and deployment are made, along with comparisons of coordinate-to-metric translation schemes for their ability to recover biomechanical measures. All these comparisons are based on reported peer-reviewed measurements, not on results from a particular benchmark suite.
- 3) **Error propagation analysis.** We find an expression for a first-order bound on landmark localization error translated into joint-angle error and demonstrate that the amplification factor is unbounded precisely at the configurations (full extension and full flexion) that contain the greatest amount of diagnostic information in sport contexts. This analysis gives insight into why reported landmark accuracy exceeds kinematic accuracy.
- 4) **A temporal validity model.** Feedback deadlines are turned into latency budgets, even accounting for the structural latency of windowed lifting models, and the model is used to differentiate between live, near-live, and retrospective usage contexts that are lumped together in the literature.
- 5) **A critical analysis of validation practice and a reference implementation.** We discover that Stage II and Stage IV are the least developed stages despite being the very stages that all practical claims hinge upon. In order to render our argument tangible and falsifiable, we provide a full implementation of the pipeline on consumer hardware with throughput figures for both CPU and GPU as well as explicit claim statements.
- 6) **Six closing gaps and reporting standard.** We define six gaps expressed as observable lacks, and outline a simple, cheap reporting checklist which will make comparisons across studies feasible without new methodology.

D. Organization

The pipeline formalization and derivation of the error propagation and latency bounds are given in Section II. The review methodology is outlined in Section III. The taxonomy is provided in Section IV. Literature review for each of the stages is presented in Section V. Comparative analysis is performed in Section VI. Reference implementation is given in Section VII. The synthesis of converging gaps is presented in Section VIII. Open challenges and agenda are stated in Section IX, followed by conclusions in Section X.

II. BACKGROUND AND PROBLEM FORMULATION

A. Human Pose Estimation as a Sensing Layer

In the context of this survey, HPE estimates a set of anatomical joint coordinates x , given an input frame f_θ . For a video of T frames, the pose sequence estimation results in

$$\hat{P} = f_\theta(x) = \{\hat{p}_{tj}\}_{j=1}^J, \quad \hat{p}_{tj} = (x_{tj}, y_{tj}, z_{tj}, v_{tj}), \quad (1)$$

where J is the number of landmarks, $\hat{p}_{tj}$ is the j -th landmark of frame t , and $v_{tj} \in [0,1]$ is a visibility or confidence value. The difference between the 2D and the 3D estimation problems is the presence or absence of recovery of z_{tj} ; the difference between the top-down and bottom-up approaches is whether a person bounding box detection occurs before joint regression [10], [11]. Heatmap-regression encoder [12], [13], direct-regression and lifting networks [8], [14], and mobile landmark regressors vary in their accuracy, latency, and number of landmarks exposed, i.e., 17 COCO keypoints, 19 plus 2 additional derived midpoints, or 33 of the BlazePose structure.



In the context of analyzing motion, v is just as important as the coordinates themselves. A landmark that is reliably detected by the network, despite being occluded by the athlete's own body part, will stealthily contaminate all further calculations, unless it is filtered out. A rule like $v_{tj} \geq \tau$, where the constant τ is set equal to the occluded/visible decision point in the network, is thus no trivial choice in terms of measuring accuracy.

B. From Landmarks to Biomechanical Quantities

Coordinates do not possess any intrinsic biomechanical information. A coordinate triple does not tell whether the knee was extended too much, too little, or unequally compared to the opposite leg. Most motion analysis software requires the addition of a transformation step for this reason. The most frequent transformation used is joint angle calculation using vector geometry, whereby a joint angle is calculated from the coordinates of points A , B and C ,

$$\theta_B = \arccos \left(\frac{(A - B) \cdot (C - B)}{\|A - B\| \|C - B\|} \right), \quad (2)$$

evaluated relative to whatever subset of the coordinate axes is deemed trustworthy. The equation for (2) is universal in applying to all types of joint—knee, hip, elbow—except for the triplet of landmarks. It is that universality that allows an analysis engine to be agnostic of the movement it analyzes.

Three additional quantities can be calculated based on (2). The left-right asymmetry is normally defined as a symmetry index

$$SI = \frac{2 |\theta_L - \theta_R|}{\theta_L + \theta_R} \times 100\%, \quad (3)$$

angular velocity follows by finite differencing, $\omega_t = (\theta_t - \theta_{t-1})/\Delta t$, and deviations from a reference range produce an aggregate score

$$S = \max \left(0, 100 - \sum_{i=1}^k w_i |\theta_i - \theta_i^{\text{ref}}| \right), \quad (4)$$

where θ_i^{ref} is the reference angle and w_i weights joint i based on its relevance for injury risk or performance metrics. Deep-learning-based pose-quality analysis [15] proposes a similar scoring function but trained directly from pose data instead of designed by hand.

The following scoping definition should be made. A model may estimate the 3D coordinates while performing all biomechanical calculations on the 2D plane projection since the depth dimension is more unreliable than other dimensions for lightweight backbones [16], [17]. Mixing the terms “3D pose estimation” and “3D biomechanical analysis” would be incorrect, and we make the distinction hereinafter.

C. Error Propagation into Kinematic Quantities

Landmark accuracy is reported in the literature as an average position error per joint or percentage of keypoints that have been correctly localized. What the coach or physical therapist cares about, on the other hand, is the angle. The transformation from one to another is non-linear, and so is the non-linearity.

Let $\mathbf{u} = \mathbf{A} - \mathbf{B}$ and $\mathbf{v} = \mathbf{C} - \mathbf{B}$, so that $\theta = \arccos(\hat{\mathbf{u}} \cdot \hat{\mathbf{v}})$. Perturbing the landmarks by $\delta \mathbf{u}$ and $\delta \mathbf{v}$ and differentiating gives, to first order,

$$|\delta \theta| \leq \frac{1}{\sin \theta} \left(\frac{\|\delta \mathbf{u}\|}{\|\mathbf{u}\|} + \frac{\|\delta \mathbf{v}\|}{\|\mathbf{v}\|} \right). \quad (5)$$

Two results are immediately obvious. First, the magnification factor $1/\sin \theta$ becomes smallest for $\theta = 90^\circ$ and approaches infinity as $\theta \rightarrow 0^\circ$ or $\theta \rightarrow 180^\circ$ (Fig. 1). The kinematic configurations in sport that are most meaningful—the nearly complete extension of the elbow joint at contact with the shuttle, the large knee bend at the bottom of the squat—are precisely those that are poorly conditioned. Second, since the error bound is scaled by segment length, a foreshortened segment will exaggerate the error: if a segment is projected at an out-of-plane angle ϕ , then it appears with length



$$\frac{\delta \ell}{\ell} = -\tan \phi \delta \phi, \quad (6)$$

with the measurement breaking down as the element approaches the optical axis. In other words, the monocular camera suffers worst with motion toward or away from the camera, a frequent occurrence in field-based sports and an uncommon one in laboratory tests.

The upshot is that, contrary to the popular notion, a 5 mm landmark error and 1° angular error do not equate as units of error; the latter cannot be deduced from the former without knowledge of the joint angles. This accounts for a trend we uncover in Section VI: despite having similar accuracy in terms of landmark errors, kinematic systems can vary widely in their agreement because they are tested in different joint angles.

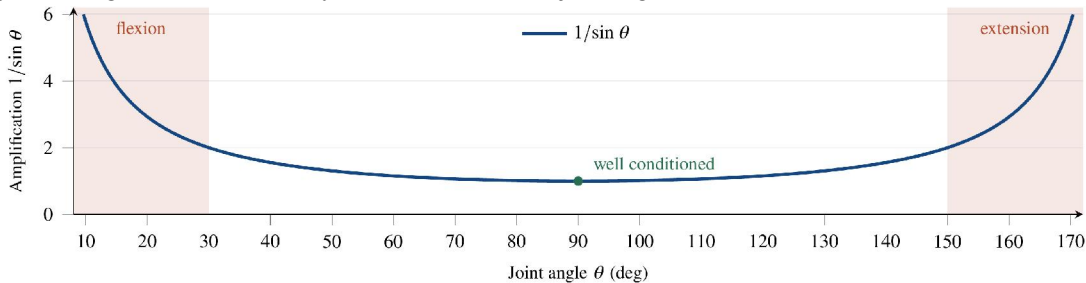


Fig. 1. First-order transformation of landmark error to joint-angle error, based on (5). $1/\sin \theta$ is one when $\theta = 90^\circ$ and tends to infinity as the joint approaches full extension and full flexion; these are indicated by the hatched areas. The athletic keyframes occur in the hatched areas, so the most informative frames are also the least conditioned. This relation is geometric and does not depend on any specific estimator.

D. Temporal Validity and the Latency Budget

A measurement which comes after the movement has concluded will have no effect upon it. It seems like a basic point which is frequently overlooked; the literature often treats a system as being "real-time" purely by its inference time per frame. The latency of the entire pipeline is

$$L_{e2e} = L_{cap} + L_{dec} + L_{inf} + L_{sm} + L_{seg} + L_{fb} + L_{ren}, \quad (7)$$

where L denotes the capture, decode, inference, temporal smoothing, segmentation, feedback generation, and rendering, respectively. For a causal estimator on per-frame basis, L_{inf} is the measured latency per frame. However, for a window-based estimator, the estimator itself is non-causal: with a window of W frames at a sampling frequency f_s , the minimum latency for a window is

$$L_{win} = \frac{W - 1}{2f_s}. \quad (8)$$

and for $f_s = 30$ Hz, with 27 frames in a window—an often-used configuration for video-based lifting [9]—the latency is 433 ms before performing any processing. This amount of latency is greater than the time required for the fastest phases of many sports activities.

The feedback deadline is defined as Δ_c , the duration in which the cue has to arrive for the movement to be influenced by it and $L_{e2e} \leq \Delta_c$ is enforced. As Δ_c spans three orders of magnitude depending on the application, the criterion is further split into three levels (see Table I). The three levels are not merely cosmetic, since a system meeting the first level can be based on an offline 3D lifter with precision, while a system meeting the live level is designed with causality and light weight in mind and makes the precision trade-offs described in Section VI.



TABLE I
TEMPORAL VALIDITY LEVELS FOR POSTURE-BASED MOVEMENT ANALYSIS

Level	Time limit and frequency	Common computation	Example application
Live	$\Delta t \leq 0.1$ s; ≥ 30 fps	edge CPU/GPU, causality-based	prompt within a set or repetition; gait correction
Near-live	$\Delta t \leq 30$ s; ≥ 1 fps	asynchronous CPU queue	assessment of set or repetition before subsequent trials
Historical	minutes to hours	GPU batch processing	evaluation, longitudinal analysis, studies

E. Scope Differences

There are three differences that appear repeatedly and should be settled upfront. *Estimation vs. Modeling*: estimation of 3D landmarks is not the same as estimation of joint kinematics in the anatomical coordinate frame, which involves skeleton modeling, segment definitions, and sometimes inverse-kinematics solution. *Agreement vs. Utility*: the system can agree with the reference system and at the same time be useless for coaches because usefulness relies on interpretability and timing. *Accuracy vs. Equity*: aggregate accuracy may hide subgroup biases, hence one accuracy measure cannot be taken as evidence of clinical fitness. We come back to each of these differences in Sections VI and IX.

III. REVIEW METHODOLOGY

A. Literature Search Approach

Potential articles were sourced using selective searches on IEEE Xplore, Scopus journals/conferences, ACM Digital Library, PubMed journals, and CVPR, ICCV, NeurIPS, WACV, and CVPRW conference proceedings. The period of search was from January 2021 to September 2026 owing to the fast-changing nature of deep learning human pose estimation algorithms but while still ensuring that the latest systematic reviews in the field of post-2020 sports-related articles remained intact. Keywords used included "human pose estimation", "3D pose estimation", "markerless motion capture", "sports biomechanics", "joint angle estimation", "action quality assessment", "athlete", and backbone models' names.

B. Inclusion/Exclusion Criteria

A paper was considered eligible if it fulfilled at least one of the following three criteria: (i) the paper suggested or compared HPE frameworks pertinent to video motion capture; (ii) the paper reviewed HPE methods in sports or exercises through systematic review or PRISMA survey; or (iii) the paper described an application platform, data set, or evaluation study of sports/clinical motion capture based on visual pose.

A total of four exclusion criteria were applied. First of all, pre-prints and papers from venues that were not peer reviewed were excluded based on the notion that any survey aimed at informing the design of new technologies should be founded on refereed materials only. It has to be mentioned that there is a certain price associated with this criterion since neither the original BlazePose paper nor MediaPipe Pose description is a refereed publication. That is why BlazePose is considered according to the peer-reviewed papers evaluating it [16]–[19]. Another exclusion criterion was the absence of biomechanical output from the algorithm unless an architecture was proposed in the paper that was further used for other sports applications. Static-pose recognition without any temporal component, doctoral dissertation, book chapter, and materials that had no DOI were also excluded.



C. Screening and Corpus Composition

Screening followed a PRISMA-informed procedure. Titles and abstracts were screened for topical relevance; full texts were retrieved for candidates addressing at least one pipeline stage; each retrieved item was assigned a single primary stage reflecting its main contribution, so that the per-stage counts in Table II are mutually exclusive and auditable. The final corpus comprises 45 peer-reviewed studies. Table II summarizes its composition and already exposes the central asymmetry of the field: the two stages that determine whether a measurement can be trusted—translation and validation—together account for 18% of the corpus, while estimation alone accounts for 42%.

TABLE II: COMPOSITION OF THE REVIEWED CORPUS ($N = 45$). STAGE ASSIGNMENT IS MUTUALLY EXCLUSIVE AND IDENTIFIES THE MAIN CONTRIBUTION OF EACH STUDY

Stage	<i>n</i>	Share	Dominant approach	Evidence type
I. Landmark estimation	19	42%	spatio-temporal transformers; lightweight CNNs	benchmark MPJPE / AP on laboratory datasets
II. Coordinate-to-metric translation	3	7%	three-point vector geometry	internal consistency; reference ranges from literature
III. Sport-specific application	14	31%	single-sport pipelines	descriptive kinematics; no external reference
IV. Validation and implementation	5	11%	agreement study vs. mocap or IMU	correlation, RMSE, occasional Bland–Altman
Cross-cutting surveys	4	9%	narrative and PRISMA synthesis	qualitative synthesis

D. Appraisal Criteria

Given the methodological diversity of the corpus, a numerical assessment score is not applicable. The articles were appraised according to five criteria, selected as they determine whether the reported numbers are interpretable:

- D1) Coordinate space.** Is the coordinate system for the computation of angles described (2D projection, 3D camera space, or anatomical frame)?
- D2) Handling of visibility.** Is there a threshold on confidence values used to exclude/interpolate/proagate low-confidence landmarks?
- D3) Provenance of the reference range.** Are the reference ranges provided, and are they population-based or individual-specific?
- D4) Design of validation.** Is the agreement with an independent reference measured, and if so, by what statistic?
- D5) Deployment information.** Is deployment information (hardware, framerate, resolution, camera geometry, latency) provided?

These aspects are intentionally kept very few in number. They relate to the disclosures that cost the author a sentence each but are those that will decide if two accuracies reported are comparable. We have used them qualitatively in Section VI and converted them into a submission checklist in Section IX.

E. Threats to the Validity of This Review

Four threats bear mention. First, coding was done by the authors themselves rather than by independent reviewers, so that staging is necessarily subjective, with the mutually exclusive criterion and the counts for each stage being provided for auditability rather than objectivity. Second, the corpus is limited to the English language, peer-reviewed literature, thus omitting a vast amount of practice in applied literature that appears in national journals and conferences; exclusion of preprints means that the first publication of the primary descriptions of several popular lightweight estimators, such as BlazePose, is missing from the literature. Third, the 2021–2026 time frame misses the pre-2021 architectural



literature, which we cite only when we need it for explanation of some mechanisms. Fourth, the variety of outcomes in the corpus makes meta-analysis impossible; hence, we describe only trends and absences, rather than meta-effect sizes, and treat the six gaps in Section VIII as claims about the absence of certain information from the corpus, rather than from the entire field.

IV. A PIPELINE-CENTERED TAXONOMY

The literature surveyed is structured according to a four-stage pipeline (Figure 2). The organizing premise here is that every stage comes with its own set of vulnerabilities, its own set of metrics, and its own level of maturity; moreover, a system's overall integrity cannot be greater than that of its weakest link. Categorization by sport would not reveal that every time the geometry of a sport is rediscovered independently; categorization by architecture would not show that architecture was a minor component in the deployed system.

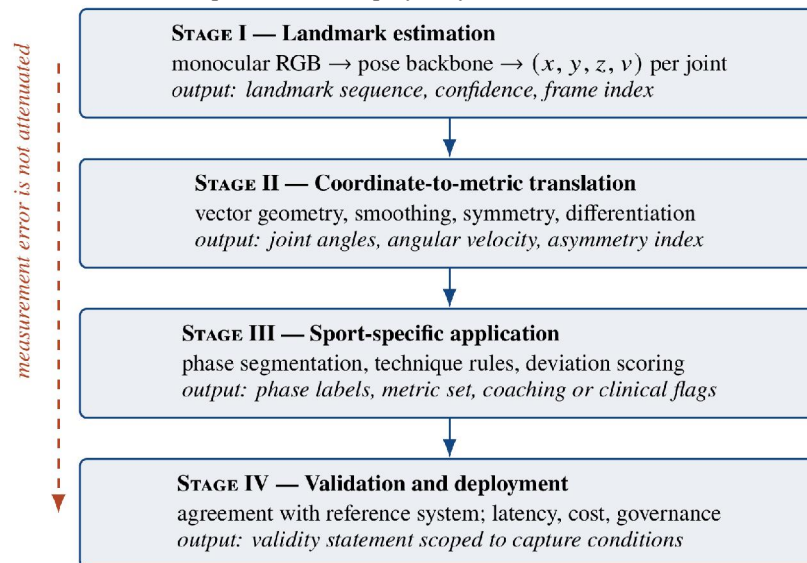


Fig. 2. The four-step estimation-to-biomechanics pipeline on which the survey framework is based. Step 1 estimates anatomical landmarks from video, Step 2 converts coordinates into biomechanical quantities, Step 3 applies them to a particular motion, Step 4 ensures consistency between the method and an independent reference and measures deployment cost. The dotted line denotes the property that drives the taxonomy: measurement errors introduced in Step 1 are not corrected by subsequent steps, and so the reliability of the results is limited by the weakest step, not the strongest one. The per-step reference counts in Table III quantify where research effort is dense, not where error is greatest.

The taxonomy is intentionally asymmetric in its reference density. Step 1 is the most dense, competitive, benchmarking-driven part of the literature, with 42% of all references. Step 2 and Step 4, on the other hand, are sparse: the translation and validation layers together comprise 18% of the literature, even though they are the two layers on which all claims rest. This asymmetry is one of the main findings, and we discuss it again in Sections VI and VIII.



TABLE III: A FOUR-STAGE TAXONOMY FOR THE ESTIMATION-TO-BIOMECHANICS PIPELINE. THE NUMBER IN PARENTHESES IS THE TOTAL NUMBER OF REVIEWED REFERENCES THAT MAKE A PRIMARY CONTRIBUTION TO EACH STAGE; FOUR CROSS-CUTTING SURVEYS ON SPORTS AND HPE ARE TREATED INDEPENDENTLY. THE STAGES ARE SEQUENTIAL AND INTERDEPENDENT

Stage	Function	Primary Failure Mode
Stage I: Landmark estimation (19)	recover key anatomical landmarks (x,y,z,v) per frame	occlusion, truncation, out-of-distribution pose, weak depth axis
Stage II: Coordinate-to-metric translation (3)	convert coordinates to joint angles, symmetry, velocity	undocumented reference ranges; plane restriction implicitly assumed
Stage III: Sport-specific application (14)	apply metrics to gait, exercise, field, or water sports	single-exercise pipeline; no generality across movements
Stage IV: Validation and deployment (5)	validate against a reference system; measure deployability	self-assessed accuracy; no independent ground truth
Cross-Cutting (4)	surveys on HPE applications in sports	scope overlap; conclusions not independent

There are two features of the taxonomy that need to be highlighted. First, the stages are *sequential*, since the output of one stage becomes the input of the next, and so it is possible for a system to excel at Stage I but be unusable at Stage III even though there is no single faulty component. Second, the stages are *coupled* through error, not just data: as shown by (5), the severity of a Stage I error is based on the combined configuration found at Stage II, and hence the movement chosen at Stage III.

A third feature is diagnostic in nature. Since each stage has its own characteristic failure mode, a prediction of the taxonomy is that the places where systems fail can be determined ahead of time. A paper that reports only Stage I results will be expected to report nothing about either Stage II disclosure or Stage IV agreement; a paper that reports Stage III results without Stage IV agreement should be expected to generalize poorly to new situations.

V. LITERATURE REVIEW

A. Stage I: Landmark Estimation Architectures

1) Spatiotemporal Transformers for Pose Lifting

The state-of-the-art in terms of accuracy for mono 3D HPE relies on transformers that leverage temporal information in order to deal with depth ambiguity and self-occlusions. Zheng et al. [8] introduced PoseFormer, an architecture composed only of transformers, which exploits spatiotemporal dependencies through self-attention, discarding the use of convolutions and LSTM used in previous video-based lifting approaches. It sets the state-of-the-art on Human3.6M and MPI-INF-3DHP while being characterized by a structural latency, measured using (8), which prevents its use in live feedback applications.

Zhang et al. [9] tackled a different type of artifact using MixSTE by changing the problem formulation to sequence-to-sequence, whereby the network returns an entire pose sequence instead of only one center frame estimate. This change of formulation leads to a decrease in temporal artifacts, particularly in cases of fast non-periodic motion, which is very common in sports. Nevertheless, the benchmarks presented in MixSTE are offline using pre-recorded sequences and do not provide any means for translating the coordinates to biomechanical quantities.

This recent literature further improves upon spatio-temporal modeling. In their approach, Du et al. [20] encode kinematic priors directly within a spatio-temporal feature transformation, explicitly encoding assumptions about the anatomy of the skeleton, rather than implicit learning from data. Meanwhile, Liu and Zhang [21] suggest STGFormer, which recasts the spatio-temporal problem in the context of a graph with nodes representing joints. Lang and Chuah



[22] develop an event-guided attention mechanism, with adaptive rather than uniform temporal aggregation. However, all of the reviewed work in this line share a common drawback: neither provides a latency budget for live training, nor do they use athletic videos.

2) *Vanilla Vision Transformers and Task-oriented Surveys*

At the same time, Xu et al. [7] demonstrated using ViTPose that vanilla vision transformers, which lack hierarchical architecture specific to HPE and do not include domain expertise in their architecture, make an effective and scalable baseline for 2D keypoint detection, reaching state-of-the-art performance on the COCO keypoint benchmark. This finding is crucial for motion analysis as it implies that domain-specific expertise embedded into the architecture is unnecessary for good performance.

This stage is represented by five surveys that examine the literature from different angles. Zheng et al. [5] study the evolution from heatmap-regression CNNs to transformer-based spatio-temporal models. They classify the area along the axes of 2D vs. 3D and single vs. multi-person methods. Liu et al. [6] survey monocular 2D and 3D estimation from the perspective of deep learning, whereas Liu et al. [23] broaden the subject matter to include mesh recovery (relevant when joint angles are not enough to describe the orientation of segments). Yan et al. [24] conduct a survey for the sake of efficiency, making the accuracy-efficiency trade-off the key axis of interest. El Kaid and Baina [25] perform a systematic review of 3D HPE methods from deep learning. None of these surveys has been conducted from the perspective of the measurement downstream.

Two architectural concepts from the pre-2021 literature retain their structural significance because they live on in subsequent architectures. The heatmap-regression encoder in the stacked hourglass architecture family [12] and its high-resolution cousin, HRNet [13], provide the benchmark of localization performance for lightweight alternatives. Direct regression from images to 3D poses in VNect [14] proved that monocular 3D pose estimation is possible in real-time without passing through a 2D intermediate stage. Bone-based pose decomposition by Chen et al. [26] revisited this approach using a decomposition of the skeleton into a kinematic tree such that estimation and classification of pose become one joint task architecturally unique in that it integrates a biomechanical decomposition within the estimator itself.

3) *Lightweight and On-Device Estimators*

Latency is the primary constraint for coaching on consumer hardware and in such scenarios, lightweight architectures take precedence. Zhu et al. [27] proposed a cross-fusion lightweight architecture designed for edge-device inference, while Lou et al. [28] sought a lightweight estimator with an adaptive regression loss that skips the computationally heavier heatmap encoding step.

The most commonly used lightweight estimator in sports applications is BlazePose, which is the pose backend of MediaPipe. In a benchmarking study of BlazePose vs. OpenPose, Mroz et al. [16] observed that BlazePose was significantly faster than OpenPose for CPU only consumer hardware, providing denser landmark set with less reliable estimates of depth axis as compared to planar coordinates. The impact of the pose backend on a downstream biomechanical quantity was shown by Mundt et al. [17], who compared AlphaPose, BlazePose, and OpenPose in terms of ground reaction force estimation based on 2D pose landmarks and showed that the selection of pose backend had a significant impact on the estimation accuracy.

The latest rigorous study in this regard is that of Kahl et al. [19], who benchmarked the monocular HPE frameworks using videos for standard tasks and found substantial variation in accuracy among different published methods, even greater than the variation in accuracies between different frameworks. We address this directly in terms of methodology in Section VI.

4) *Multi-Person and Occlusion Robust Estimation*

Single athlete approaches have assumed one person per video, which is not a good assumption for team sports and exercises. Multi-person approaches have evolved from part affinity field based bottom-up approach [10] and region-based top-down approach with tracking [11] to occlusion handling [29] and dual graph networks for crowded scenes [30]. The latter two methods show significant improvement in highly occluded scenarios.



For the task of movement analysis, on the other hand, these techniques solve an entirely different problem, which arises when trying to isolate several moving subjects, but not the one that involves surviving self-occlusions caused by the athlete's own limbs in extremely wide ranges of movement. The sports literature that was surveyed pays much more attention to the latter problem than the former one, and the two problems are different – crowd-occlusion challenges are not a good way to test self-occlusion regimes in a tennis serve or a gymnastics exercise.

B. Stage II: Pose Estimation to Biomechanical Quantities

It is at this stage that pose estimation becomes biomechanics, and this stage is performed most often informally. Joint angles estimation under real-world conditions, in which 3D HPE was applied to the evaluation of angles, was considered by Choi et al. [31], and the joint angle computation approach was provided by Ciardiello et al. [32] as a pose estimation extension via deep learning.

There are two conclusions one can draw. First, (2) is universally adopted, thereby showing that the translation part is already done mathematically; the point of contention lies in the choice of coordinate axes and visibility considerations, not in geometry itself. Second, the ranges of angles used in (4), θ_i^{ref} , are almost exclusively based on population-level biomechanics literature, not on individual calibration for the specific athlete; hence, the performance score measures deviation from a population-level standard rather than an individual one. It is a serious limitation, not a cosmetic one, as it constrains interpretation of the score meaning.

This point about plane restriction is especially important. Due to the reduced depth-axis contribution of lightweight structures [16], [17], it makes sense to restrict the computation of angle to the 2D projection on the basis of considerations of accuracy preservation. In addition, it is not always stated, meaning that the reader cannot tell whether a system reports the true 3D angles or simply the 2D projection. As (5) and (6) show, there is no reason why these should be approximations of each other; rather, the 2D version is distorted by foreshortening in an unpredictable manner based on the relative position of camera and athlete, and most of all when the motion contains the most information.

C. Stage III: Sport-Specific Application

1) Gait and Locomotion

Gait is an area of application that has received the most development both through clinical and sporting applications requiring economical evaluation methods. Stenum et al. [33] have shown 2D gait analysis via pose estimation using video, followed by extension to clinical cohorts and longitudinal evaluation [34], highlighting the influence of viewpoint choice and camera setup on usable outputs.

Rigoni et al. [35] made significant progress in the field by aligning body pose estimation with the classic gait analysis paradigms using 3D mesh modeling, pointing to a promising future path – rather than replacing the already well-defined biomechanical parameters with their approximations from the poses, including the existing frameworks in body pose estimation. Concerning running in particular, Li et al. [36] compared wearable inertial sensors to optical motion capture technology for the purpose of lower limb running biomechanics.

This additional warning comes from the clinical literature and is one that is frequently overlooked in the sports literature. According to Wade et al. [37], there is a significant disparity between the accuracy of laboratory measurements and markerless gait analysis, which differs depending on the joint and plane: It is small for sagittal plane hip and knee joint angles while walking slowly and significantly greater for quantities in the frontal and transverse planes, namely those involved in many injury mechanisms.

2) Strength, Resistance, and Field Movements

Squats and related resistance exercises dominate because they are reproducible, cyclic, and easy to stage in a single viewpoint. Haberkamp et al. [38] validated an AI HPE model for measuring single-leg squat kinematics, one of the few reviewed studies to report validity coefficients against a reference system for a specific resistance movement.

The field sport problem entails an entirely new challenge due to the rapid motion and coarse resolution of the subjects in an unstructured scene. The problem of metric reconstruction of 3D pose in a single sports broadcast video was solved by Baumgartner and Klatt [39] using partial registration of the sports field, without requiring multi-camera



calibration which is the limitation of broadcasting analysis, and furthered by Baumgartner et al. [40], who used track-and-field broadcasts to acquire spatial knowledge.

Motor-sports and multi-view cases have been considered by Ølstad et al. [41], who performed 3D athlete and rig pose estimation from 360-degree video in Olympic sailing, demonstrating that when there is a tight coupling between the platform and the athlete, estimating kinematics of the human body apart from the platform becomes another issue.

Aquatic case offers a completely different scenario. Coifman et al. [42] have created the pose estimation for automated biomechanical swimming analysis, where the person is occluded continuously by water, position is distorted by refraction, and movement is non-periodic. It is important to remember that the same 33-landmark classification has to be understood differently in this context.

3) *Movement Dynamics and Action Quality Evaluation*

Evaluation of technique execution beyond pure motion kinematics is addressed comprehensively in Zhang et al. [15], where the current state of deep learning techniques for action quality evaluation is reviewed in depth. The critical evaluation of the field can be found in Glazier [43], who presents the concept of markerless motion capture in sports as a compromise rather than as an absolute improvement, often understated in applied research. Glazier's opinion is consistent with the error analysis described in Section II-C, as it is not just the issue of accuracy vs. convenience but the problem of accuracy vs. configurations that can be captured.

D. *Stage IV: Validation and Deployment*

The part of the pipeline which is least served systematically by validation is this one. Validating requires having a reference system. The two ways that can be followed include comparing against a marker-based optical motion capture, and comparing against manual annotation from intra- or inter-raters. Both are costly, and both are often neglected in favor of self-reported accuracy.

Aleksić et al. [44] validated automated countermovement vertical jump analysis by comparing markerless pose estimation with a 3D marker-based motion capture system, and Ksoll et al. [45] conducted a pilot study of comparing a CNN-based markerless system with a marker-based 3D motion analysis system for gait analysis. Dill et al. [18] compared MediaPipe Pose-based 3D reconstruction using stereo camera information fusion for physical exercises, and further investigated accuracy assessment for physiotherapy exercises.

Uhlrich et al. [46] advanced the reference standard itself with OpenCap, computing both kinematics and dynamics from two or more smartphone videos, reported as roughly $25 \times$ faster than a laboratory-based approach at under 1% of the cost. Miller et al. [47] subsequently integrated machine learning with musculoskeletal simulation to improve OpenCap-based dynamics estimation. OpenCap's significance for this survey is that it establishes a credible low-cost reference against which vision-based systems could be validated at scale. That very few reviewed studies actually use such a reference is therefore a conspicuous omission.

Edriss et al. [4] supply the deployment perspective absent from architectural work, reviewing commercial vision sensors and the practical setup choices—camera placement, lighting, frame rate—that determine real-world accuracy independently of model choice. Their emphasis aligns with the temporal-validity model of Section II-D: deployment conditions and timing constraints, not architecture, determine whether a system can be used at all in a given tier.

Finally, the reporting of agreement is inconsistent. The proper statistics for method comparison are clear—Bland-Altman limits of agreement for bias and random errors [48] and intraclass correlation coefficients with clearly defined models and forms [49]; however, in the reviewed validation studies, correlation coefficients have been reported, not the statistics for method comparison, as they are only measures of association and are not sensitive to bias. A method may correlate with $r = 0.95$ but overestimate every angle of knee flexion by 8° ; these two facts imply opposing consequences for application in practice.



VI. COMPARATIVE ANALYSIS

A. Stage I: Landmark-Estimation Backbone Families

The backbone families utilized in the literature on sports are listed in Table IV based on the joint criteria of suitability for motion analysis, which include localization accuracy, real-time capability, coordinate-space properties, and the stated target of deployment by the authors themselves. The qualitative data is obtained from peer-reviewed measurements; the "Design Intent" column gives the stated target and not a redetermined value of throughput since the hardware used, resolution, and batch sizes differ across works and make redetermination unreliable.

TABLE IV: SAMPLED BACKBONE FAMILIES FOR LANDMARK ESTIMATION AS USED IN THE REVIEWED MOVEMENT ANALYSIS LITERATURE. ACCURACY IS RANKED QUALITATIVELY AS IT CANNOT BE REDETERMINED GIVEN DIFFERING CONDITIONS.

Family	Representative	Landmarks	Depth output	Real-time on CPU	Design intent
Heatmap CNN	Stacked Hourglass [12], HRNet [13]	17	none (2D)	no	high localization accuracy; offline analysis
Direct 3D regression	VNect [14]	21	weak, relative	marginal	single-camera 3D without a 2D stage
Spatio-temporal transformer	PoseFormer [8], MixSTE [9]	17	relative, strong	no	benchmark accuracy; windowed latency
Plain vision transformer	ViTPose [7]	17	none (2D)	no	scalable, transferable baseline
Lightweight CNN	BlazePose / MediaPipe [16], [17]	33	weak, relative	yes	on-device, wearable-free deployment
Lightweight transformer	LAR-Pose [28]	17	none (2D)	yes	efficiency-constrained inference

Firstly, only lightweight architectures are capable of real-time throughput in CPU-only settings, and thus they represent the sole realistic candidate for live coaching feedback in the live tier of Table I. Secondly, this advantage comes at the price of reduced reliability in the depth axis, which needs to be explicitly handled at Stage II. The findings of Mundt et al. [17] show that the impact of this price is not just theoretical, but very concrete indeed: selecting between BlazePose, AlphaPose and OpenPose resulted in differences in the ground-reaction forces calculated using pose estimates.

Secondly, Kahl et al. [19] introduce a methodological caveat that should be highlighted. In case accuracy significantly differs between different implementations of nominally similar frameworks, the accuracy figures cited in the papers of applied literature cannot be compared to each other. This invalidates the approach of ranking the systems based on their reported accuracy only. This also means that the implementation in addition to the architecture represents an experimental variable to be reported, and this is impossible using the methodology of Table IV since the implementations are not often made publicly available.

B. Stage II: Coordinate to Metric Transformation

Table V below reveals that no single approach typifies the reviewed literature; instead, systems make up these approaches. The mathematical framework for the translation module is set. What remains to be determined are calibration (to what range of values, individually tailored or average) and disclosure (of what coordinate system,



explicit or implicit). This table also explains the sparseness of Stage II in the literature, since each individual approach involves only a few lines of code and hence is not publishable in itself, despite the fact that the decision made in each such few lines defines the value of every subsequent number.

TABLE V: COORDINATE TO BIOMECHANICAL TRANSLATION APPROACHES USED IN THE REVIEWED LITERATURE, ALONG WITH THE QUANTITIES THEY CAN RECOVER AND THEIR LIMITATIONS

Strategy	Quantities recovered	Principal limitation
Three-point vector geometry	joint angles in a chosen plane	out-of-plane deviations undetectable; ill-conditioned at full extension
Symmetry index	left-right asymmetry	requires both limbs visible; normalizes away absolute magnitude
Temporal phase detection	repetition count, cycle boundaries	threshold-dependent; cyclic movements only
Weighted deviation score	aggregate performance score	reference ranges rarely individualized; weights unvalidated
Learned action quality assessment	technique score without explicit angles	interpretability reduced; data hungry

C. Stage III: Dataset Availability

The dataset situation directly determines which sports can be studied. SportsPose [50] provides dynamic 3D capture of fast, high-range-of-motion athletic movements, addressing the imbalance Uçar et al. [3] identified. More recently, AthletePose3D [51] introduced a benchmark explicitly coupling 3D pose estimation with kinematic validation in athletic movements, and AthleticsPose [52] provides authentic field-recorded sports motion with an evaluation of monocular 3D pose estimation ability on it.

The design of AthletePose3D is methodologically important in that it considers the ground truth of kinematics to be part of the benchmark itself, not an additional comparison done afterwards, which is the right way to do things at Stage IV. General-purpose benchmarks like Human3.6M and MPI-INF-3DHP, which are used throughout Stage I, have been designed in the context of indoor activity and do not reflect athletic movement; therefore, the accuracy of any results based on these datasets can only go so far when considering sports.

A gap still exists even among the new generation of datasets. Each of the three datasets supplies the ground truth landmarks and joint locations for a fixed set of activities; however, none of them gives the *reference range* that (4) asks for, or the phase transition definitions required for Stage III. That is to say, the datasets provide Stages I and some of IV ground truth, whereas Stages II and III ground truths are simply conventions.

D. Stage IV: Validation Approaches

As Table VI demonstrates, there is an important difference between claims made in the applied literature and not made explicit enough, which is the difference between measurement agreement with the reference system and the practicality of use by a coach or clinician. Aleksić et al. [44] and Ksoll et al. [45] provide the former, but not the latter, as the latter would require an experimental setup involving expert raters and usually longitudinal results. This is the most significant methodological issue identified through this survey.

TABLE VI: VALIDATION APPROACHES IN THE REVIEWED LITERATURE AGAINST THE REFERENCE STANDARD EACH DOES AND DOES NOT ESTABLISH

Reference standard	Establishes	Does not establish
Optical motion capture	joint-level agreement; temporal accuracy	ecological validity; coach usability



Reference standard	Establishes	Does not establish
Inertial measurement units	lower-limb agreement; field portability	upper-limb and planar angles
Intra- and inter-rater annotation	repeatability; agreement with expert annotation	absolute kinematic accuracy
Stereo/multi-camera fusion	depth accuracy of monocular output	single-camera usage
Self-reported accuracy	implementation-level feasibility	virtually nothing independent

First, the reference standards are not interchangeable: the optical and the inertial unit systems of motion capture show disagreement between themselves in the transverse plane; therefore, validation to one of the reference systems does not mean validation to another one. Second, “agreement” remains undefined if it is not stated which statistic is referred to, as mentioned in Section V, correlation is not agreement [48], [49]. Third, self-reporting accuracy is not a less effective form of validation but its complete absence, and it should be stated in this way.

E. Cross-Cutting Finding: Accuracy Is Not the Binding Constraint

All three tables lead to a counterintuitive conclusion when combined. Accuracy is the most optimized and the most reported parameter of the pipeline but is not the factor preventing vision-based analysis of movements from wide usage. The factors constraining are the following: lack of standardization and documentation of the translation layer (Section VI); insufficient validation against independent reference standards; and specification of the deployment parameters (camera geometry, lighting, frame rate) influencing practical accuracy more than the model used [4].

This conclusion can be used to draw a positive implication. If accuracy is binding, then progress must come from improvements in model development, which is costly and uncertain. But since accuracy does not matter and the binding requirements are documentation, agreement statistics, and capture specification, then progress is cheap. A system that reports somewhat lower accuracy for landmarks but includes the necessary documentation, agreement statistics, and a movement-general translation layer will be much more valuable to a practitioner than a system that reports higher accuracy without all these properties.

F. Positioning Relative to Prior Surveys

In Table VII, this review is contrasted with the four cross-cutting surveys in the corpus. While the distinction may seem arbitrary, prior reviews organize in terms of method families, datasets, or commercial sensors; this review is organized in terms of location in the measurement chain and therefore places equal importance on Stages II and IV.

In other words, the difference is one of axis and not of content.

TABLE VII
POSITIONING RELATIVE TO THE CROSS-CUTTING SURVEYS IN THE CORPUS

Survey	Scope	Organizing axis	Treatment of Stage II/Stage I V	Distinction of this review
Badiola-Bengoa and Mendez-Zorrilla [1]	camera-based HPE in sport and exercise (to 2021)	application area	noted as limitations	adds post-transformer literature and a formal error analysis



Survey	Scope	Organizing axis	Treatment of Stage II/Stage I V	Distinction of this review
Aulton et al. [2]	deep-learning HPE in sport	sport and task	validation flagged, not analysed	treats agreement statistics as an appraisal dimension
Uçar et al. [3]	pose estimation and tracking in sports	methodology and datasets	dataset-centric	adds deployment tiers and latency budgets
Edriss et al. [4]	commercial sensors and AI frameworks	sensor technology	deployment-centric	adds the translation layer and kinematic conditioning
This survey	45 studies, 2021–2026	position in the measurement chain	structural object of study	pipeline taxonomy, error propagation, reference implementation

VII. REFERENCE IMPLEMENTATION

A. Why Report an Implementation in a Survey

The six gaps that are combined in Section VIII are statements on what is missing from the corpus. Such statements are much more helpful if there is some physical object to test them against, and design decisions involved in such statements—dual-tier inference, a translation layer agnostic with respect to movements, and claims boundaries—become tangible when implemented, not when just described. For this reason, we provide an implementation of the whole pipeline used in the study. The implementation is a research prototype, not a commercial-grade solution; in the context of this study, it serves to demonstrate the engineering implications of the design considerations outlined in Sections II-C and II-D.

B. Architecture

The system is composed of five layers (Fig. 3), which have explicit contracts between them: presentation layer (React, Vite, Tailwind CSS), API and services layer (FastAPI, Uvicorn, Pydantic), AI engine layer (YOLOv8-pose and MediaPipe adapters over OpenCV and NumPy), sport-specific analytics layer accessible through a registry, which maps a sport name to a pipeline and an implementation of analytics, and finally, the output layer, which stores JSON artifacts and an annotated MP4.

The following two design decisions can be directly derived from Section VI. First, the translation layer is shared rather than replicated for each sport because (2) is true universally. Secondly, the pose backend is an adapter to a shared schema and thus enables a fair comparison of the two estimators using identical video material without modifying the downstream processing—without which a comparison, as shown by Mundt et al. [17], is not possible.



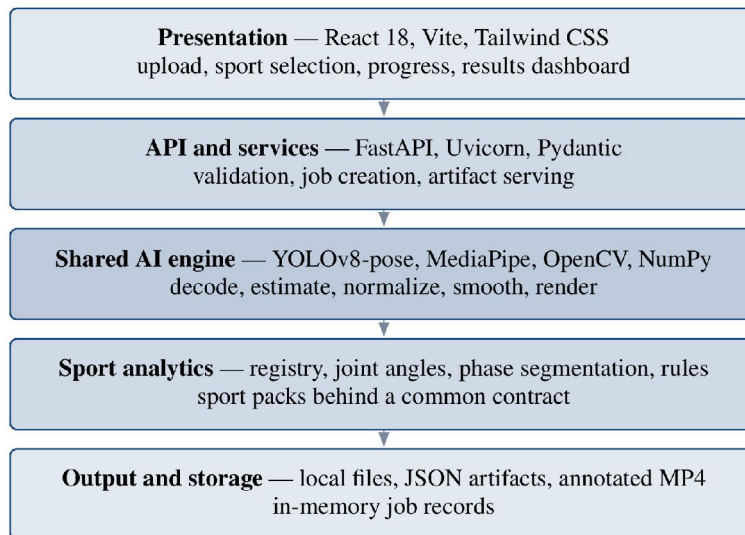


Fig. 3. Layered architecture of the reference implementation. The AI engine is shared among sports, whereas the pose backend is an adapter and hence the estimators can be swapped without affecting either the translation or the analytics layers. The registry maps a sport to a pipeline and an analytics contract; this is how the specific knowledge of movements is encoded into the configuration.

C. Two-Tier Inference and Data Flow

This particular implementation involves two estimators that have complementing working modes. First, there is a 33-landmark lightweight CNN (MediaPipe BlazePose) used as a low-latency interactive layer, while the second one (19-landmark detector) is used as a high-fidelity layer (YOLOv8-pose, 17 COCO keypoints plus two midpoints). Due to the fact that the latter is slow on the CPU, all the calculations are carried out asynchronously by the background worker; the client gets the job ID and periodically checks its status (see Fig. 4). The data contract on each step is immutable (see Table VIII); this is the reason why it is possible to get the annotated video, the raw pose stream, and the calculated metrics separately.



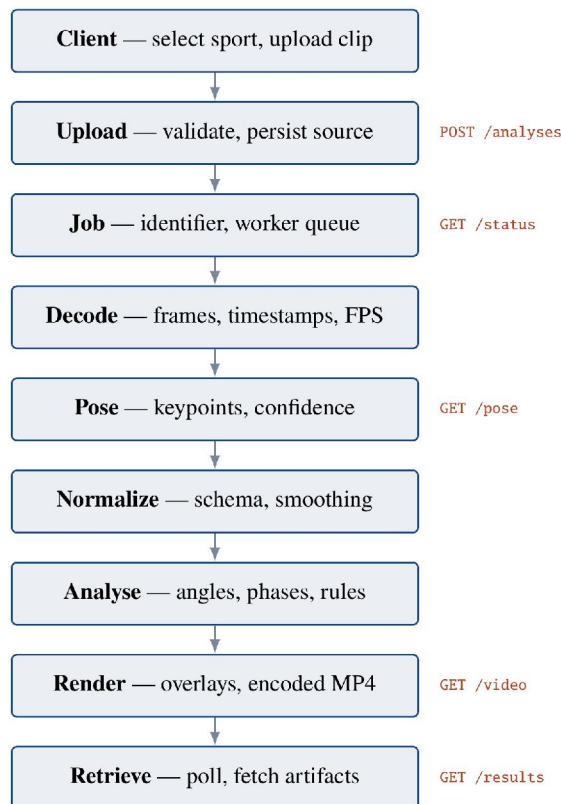


Fig. 4. Reference implementation end-to-end data flow along with API surface at each step. Asynchronous dispatch is not an optimization technique but a result of the latency analysis; on the CPU-only hardware, the high-fidelity estimator fails to comply with the live-tier deadline and, thus, the design is built specifically for the near-live tier.

TABLE VIII

STAGE TO COMPONENT MAPPING AND ARTIFACTS OF THE REFERENCE IMPLEMENTATION

Stage	Component and artifacts created
Stage I	AI engine (decode, estimate, normalize) → pose JSON: keypoints, confidence, frame index, timestamps
Stage II	landmark processor (smoothing, three point angles, symmetry, differentiation) → angle series, angular velocity, asymmetry index per frame
Stage III	sport pack and rule evaluator → phase labels, metrics, findings and coaching cards
Stage IV	benchmark harness and renderer → latency and throughput logs, annotated MP4, agreement table

D. Measurement of Performance

Performance measurements were taken for two configurations. The CPU configuration consisted of an Intel Core host with 8 GB RAM and Windows 10, along with CPU-only PyTorch; the GPU configuration included an NVIDIA RTX hardware with CUDA FP16 and TensorRT. Input was provided in the form of 720p video of a badminton overhead clear, consisting of clips from 50 to 300 frames with first five frames discarded for warming up, and measurements of mean and 95th-percentile per-frame latency recorded (Table IX and Fig. 5).



TABLE IX

Performance of the Two Estimators on the Same Footage (Measured Per Frame). MP: MediaPipe BlazePose (33 Landmarks); YOLO: YOLOv8-Pose (19 Landmarks). CPU Executions: PyTorch FP32; GPU Executions: CUDA FP16 + TensorRT

Metric	MP CPU	YOLO CPU	MP GPU	YOLO GPU
Latency (mean ms)	34.20	226.38	13.50	9.80
Latency (95th %)	48.60	254.48	18.20	14.20
Throughput (fps)	29.2	4.4	74.1	102.0
Detection success rate (%)	96.7	100.0	98.1	99.4
Confidence (mean)	0.865	0.939	0.892	0.948
Landmarks	33	19	33	19

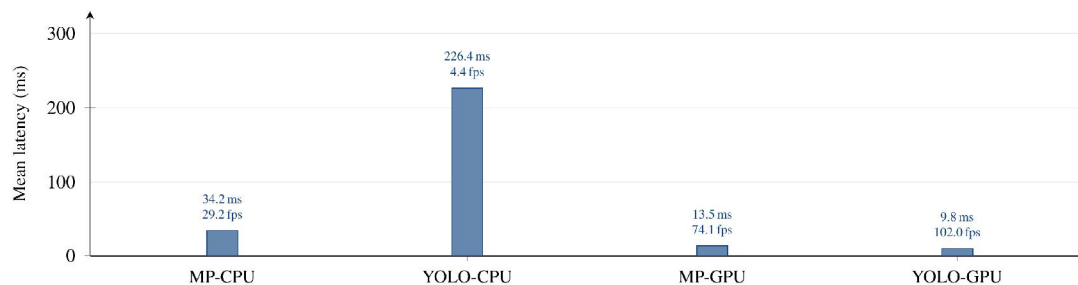


Fig. 5. Latency and implied frame rate of the four combinations of Table IX. Contrary to what was stated previously, the two estimators do not make up a ranking, but operate at two distinct performance levels. In particular, the lightweight 33-landmark model is the only one that can run interactively on CPU, while the 19-landmark detector is faster only when accelerated by a GPU. Bars start from zero.

These three results are caveats as much as findings. First, the CPU operating point of the high fidelity tier is 4.4 fps, which puts it clearly into the near-live tier of Table I: 100 frames takes about 28 s (effective 3.6 fps when accounting for decode, analytics, and rendering), which is okay between but not within repetitions. The lightweight tier, running at 29.2 fps on the same CPU, is on the border between near-live and live tiers, and this is the only one of this study's configurations that is.

Second, the detection rate is in reversed order: the detector achieves 100.0% valid detection on CPU and 99.4% on GPU, while the lightweight model 96.7% and 98.1%, respectively. This is a property of the video footage and the acceptability criterion, rather than of the hardware itself; an ordering reversal of this sort is better understood as indicating that the two sets of experiments did not use exactly the same set of clips and acceptance thresholds, as Kahl et al. [19] note. It is not an indication of superiority of CPU-based inference to GPU-based inference.

Third, the latency ratio of the two estimators on CPU ($226.38/34.20 \approx 6.6$) is much greater than on GPU ($9.80/13.50 \approx 0.7$). This is a claim about the size of the model, resolution of the inputs, and overhead in the framework on the host; it does not say anything about superior architecture. The reporting of a single "FPS" number without hardware specification, precision, and input would miss exactly this point, which is why Table IX includes all three numbers.

E. Agreement Against a Coach-Defined Reference

Four kinematic measures for the badminton overhead clear were extracted and compared against the coach-defined reference bands for the stroke (Table X). Two of the four – contact elbow angle and cocking elbow angle – have defined bands in the coaching materials used to configure the rules, and both fall within their bands. The other two are reported descriptively because no band was specified for them in the coaching materials.



TABLE X

KINEMATIC MEASURES EXTRACTED FOR THE BADMINTON OVERHEAD CLEAR AGAINST COACH-DEFINED REFERENCE BANDS. BANDS ARE COACH CONVENTIONS, NOT MOTION-CAPTURE GROUND TRUTH; TWO OF FOUR MEASURES HAVE NO DEFINED BAND

Quantity	Measured	Reference band	Interpretation
Contact elbow angle	179.5°	175–185°	inside band
Cocking elbow angle	46.8°	40–55°	inside band
Shoulder elevation	76.2°	not stated	descriptive only
Forward trunk tilt	26.3°	not stated	descriptive only

However, there is always a temptation to turn this into a claim of accuracy, as the documentation for this project originally did, namely a sub-5% kinematic error against coaching reference standards. But we shall refrain from making such a claim for three reasons immediately suggested by Section VI. First, a coaching band is a very large range, and falling within it bounds the error only at the width of the band— $\pm 5^\circ$ for contact, $\pm 7.5^\circ$ for cocking—less precise than the claim suggests. Second, since no motion-capture ground truth was captured, agreement with a convention does not necessarily represent agreement with the ground truth. Third, since two of the four quantities do not have a coaching band, then the claim cannot be made about half of the reported kinematics. The statement that can be confidently made is this: the pipeline has recovered values consistent with coaching conventions for two kinematics, and has produced plausible descriptive values for two others.

F. Claims Boundary

This case study demonstrates four things. First, it demonstrates that the entire pipeline—decoding, dual-tier estimation, normalization, angle calculation, phase segmentation, rule evaluation, and rendering—can be put together using commodity parts and run on consumer hardware. Second, it demonstrates that the two estimators exist at complementary tiers, thus making the tier terminology of Table I more than rhetoric. Third, it demonstrates that a common translation layer can service several sports by configuration, which is how it is suggested for G3. And fourth, it demonstrates that the reporting standard of Section IX can be met at negligible cost.

What is not demonstrated by this case study? Five gaps, each one corresponding to the gaps identified in Section VIII. Kinematic accuracy to an independent reference is not demonstrated, since no motion capture test was performed (G5). Generalization is not demonstrated, since only one motion was used (G3, G4). Individualization is not demonstrated, since the reference angles are population-wide coaching standards (G6). Robustness with respect to light conditions, viewpoints, skin color, body shape, and occlusions is not demonstrated since none of those varied systematically. And it does not demonstrate any clinical or coaching applicability, since the performance of the practitioner was not measured (Section VI). In this context, stating the limits of the results so explicitly is more useful than simply reporting a larger value with a more specific qualification.

VIII. CONVERGENT LITERATURE GAPS

From all four levels together, the literature reveals six gaps. These are expressed as testable absences and not as topics requiring further investigation, since an unfalsifiable gap is one that cannot be acted upon; for each gap we note what would close it.

G1 — The tension between accuracy and deployability endures. Pose-lifting based on transformer architectures currently lead accuracy [8], [9], yet are ill-equipped to provide low-latency feedback: (8) renders their latency a function of window size, not hardware speed, and therefore cannot be mitigated by faster GPUs. On the other hand, the



lighter weight architectures that enable deployability deliver less accurate depth estimates [16]. None of the reviewed systems manages to combine good geometric accuracy, depth estimation and throughput that is sustainable on a CPU.

G2 — Missing standardized coordinate-to-metric translation layer. There is no implementation or benchmark, let alone a consensus method, for translating landmark sequences into biomechanical metrics [31], [32]. While (2) has been used universally, though with unstated differences in coordinate plane, visibility treatment, and reference range, the lack of comparability even among studies using the same pose backend is evidenced by the inter-implementation differences observed in [19]. The problem could be solved with a short, version-controlled, public specification for the translation layer accompanied by a conformance test suite, like the one numeric libraries use.

G3 — Fragmentation remains movement-specific. Single-exercise workflows continue to dominate, independently verified by four reviews from 2021–2025 [1]–[4]. Extending an existing system to support a new movement generally involves re-designing the system rather than re-configuring it. This gap can be closed by showing that a unified analytical engine, tuned with a configuration file, generates the known kinematics of multiple heterogeneous movements, including a non-periodic movement, such as the throwing release.

G4 — Unbalanced datasets still exist. Team sports datasets prevail in the literature [3], despite SportsPose [50], AthletePose3D [51], and AthleticsPose [52] starting to close the gap. Generic benchmarks continue to fail representing athletic movement, and there is no existing dataset providing reference ranges and phase information for the two stages (Section VI). The gap will be closed by a dataset including ground truth from the capture level with reference ranges and phase boundaries, defined by experts, stratified by sport, sex, age, and body habitus.

G5 — Validation with reference systems is lacking. Out of 45 surveyed papers, a handful validates its work using motion capture or an independent annotation protocol [36], [44], [45]. Self-reporting of accuracy and measurement agreement with a reference system is often used; however, it does not mean that it proves practical usefulness of the system. The gap will be closed by a paper that provides limits of agreement and intraclass correlation as defined [48], [49], validated against motion capture, stratified by joint and plane, and a practical utility study as well.

G6 — The reference ranges for metrics come from the biomechanical literature on populations, not from individual athletes. Deviation scores are computed in relation to values in biomechanics of the population literature and not in relation to personalized anthropometric data, setting the bounds for the meaning of the performance score: the score shows deviation from the norm, not deviation from the pattern of this athlete. The complementary observation that the practical problems with markerless systems are significantly understated by Glazier [43] follows. A longitudinal study design that provides an estimate of individual baseline and the repeatability of deviation from this baseline would resolve the gap; this is the very quantity that an application needs.

In sum, G1 through G6 combine to describe a single unfulfilled requirement—one that can be met neither by a single one of the surveyed markerless systems nor by technological progress alone: it describes the need for a real-time-deployable, documented, movement-general, and independently validated framework for deriving biomechanically interpretable quantities from video.

IX. OPEN CHALLENGES AND DIRECTIONS FOR FUTURE RESEARCH

A. Depth Axis Stability

The most significant technical challenge is that the real-time backbones have unreliable depth information [16], [17]. There are three strategies that can be seen from the reviewed papers: limiting processing to only planar projection, which is correct but means loss of out-of-plane information such as frontal plane knee valgus; combining another view, which, as Dill et al. [18] show, works well but loses the single-camera advantage; and using prior knowledge of kinematics or anatomy, as done in bone-based decomposition [26] and kinematics-informed spatio-temporal features [20].

There is a fourth possible strategy that is not explored in the reviewed literature: reporting the calibrated depth *uncertainty*, not depth coordinates themselves, and propagating this uncertainty through (5) to get confidence intervals for the angles. The propagation is computationally cheap because the amplification factor is known analytically, and it



gives the rule engine the ability to refuse to make a decision in case the interval is too wide—a necessary feature for clinical and coaching systems that none of the reviewed systems provide. A learned depth-correction approach trained against planar-consistency constraints, with uncertainty calibrated on athletic video, is the natural next step.

B. Domain-Specific Benchmarking

The field lacks an athletic analogue of Human3.6M for the pipeline as a whole. AthletePose3D [51] is the closest existing structure, because it couples pose estimation with kinematic validation, and extending this to Stage II—evaluating biomechanical quantities rather than only landmarks—would substantially improve comparability. Two design choices would matter. First, ground truth should include phase boundaries, not only joint centers, so that Stage III can be evaluated. Second, results should be reported per joint and per plane, because the aggregate error hides the clinically decisive variation documented by Wade et al. [37].

C. Movement Generalization

All problems uncovered in G3 have the same root cause—encoding knowledge about specific movements in terms of reference angle range, phase transition thresholds, landmark subsets—in terms of configuring an identical analytical engine, instead of as bespoke software. No literature reviewed contains any system capable of doing so for a variety of different movements, let alone those that are non-cyclical, like a throw release in one sport. The example given in Section VII shows how this is done for one particular sport; the question is whether the configuration space is general enough for a movement without a cyclical phase structure. The only reasonable approach seems to be event-based segmentation, where phases are defined based on kinematic events, and not thresholds.

D. Personalized Calibration

Using athlete-specific values as reference angles and ranges, based on measurements of anthropometry, joint range of motion, or longitudinal baseline, would turn a deviation score into “deviation from an athlete’s own baseline,” instead of “deviation from a norm” [43]. This is just as much a data and longitudinal study challenge as it is an algorithmic one because what is being asked for is not agreement between the device reading and a laboratory standard but repeatability of the same measurement on the same individual, which is a specific intraclass-correlation question [49] and not a correlation question.

E. Ecological and Clinical Validation

What is needed is not simply proof of agreement in measurements, but experimental designs to validate the usability of the tool for the coach and clinician and their decision utility. In this context, the current set of validation studies [44], [45] have provided evidence of the former; not one of them provides any evidence of the latter. A minimum design would entail comparing the decision-making process both with and without the tool for the same case studies, where practitioners remain blinded to the condition, reporting only on agreement of decisions, not agreement in angles.

F. Deployment, Privacy, and Governance

Identification and privacy protections are issues for any form of identification and video, respectively, making the deployment tier of Table I have both a latency issue and a governance issue: On-device inference in the live tier implies no recording is stored anywhere except on the device, and this greatly simplifies consent and data protection compared with server-side inference. The reviewed literature is completely silent on this topic, yet this is crucial for adoption in educational, clinical, and community settings. A deployment report must state where inference happens, what is retained, and for how long, and what kind of consent it requires.

G. Equity and Subgroup Robustness

Average performance does not address biases within subgroups. Body shape, coloration, clothing, and assistive technologies impact landmark detection and a model that is unbiased on average but biased in terms of a particular subgroup will increase rather than decrease health disparities—an issue with clear clinical implications. This issue is completely unaddressed in the reviewed corpus; the bare minimum is to analyze the performance of the algorithm in the subgroup-specific manner where metadata permits and lack of this information should be treated as such.



H. A Minimum Standard for Reporting

A few simple and specific disclosures can make a big difference in comparability. Table XI provides those disclosures in the form of a checklist. Not one of them requires an experiment; each just requires one or two sentences, and in some cases, one statistic. We view this checklist as the most cost-effective approach to take short of methodological innovation, since it attacks both G2 and G5, and costs nothing in terms of review.

TABLE XI

RECOMMENDED MINIMUM REPORTING CHECKLIST FOR POSE-BASED MOVEMENT ANALYSIS

ID	Disclosure	Why it alters interpretation
R1	coordinate plane and axes for each angle	planar vs 3D angles are not equivalent; foreshortening bias differs
R2	visibility criteria and missing landmarks handling	determines how occlusions are handled: discarded, interpolated, or propagated
R3	reference ranges and their source	population average versus individual-specific baseline
R4	capture setup: camera, distance, height, frame rate, lighting	influences practical accuracy more than modeling [4]
R5	hardware, precision, input resolution, warm-up, and P95 latency	allows throughput comparisons and stratification into tiers
R6	reference system and agreement statistics (bias, limits of agreement, ICC model)	correlation $\neq$ agreement [48], [49]
R7	sport, movement, number of subjects, and composition of subgroups	defines scope and reveals unstratified error
R8	code, weighting, and configuration availability	mitigates inter-implementation variance [19]

I. Prioritized Research Agenda

Table XII ranks the actions suggested by G1–G6 according to their relative cost vs. expected benefits. The two least costly suggestions—a layer definition and use of checklists—are also those that address the constraints enumerated in Section VI, the pragmatic form of the argument of this survey.



TABLE XII

PRIORITIZED RESEARCH AGENDA BASED ON THE SIX GAPS. EFFORT IS MEASURED AGAINST A SINGLE RESEARCH GROUP; SUCCESS CRITERIA ARE PROVIDED SUCH THAT ACHIEVEMENT IS VERIFIABLE

Priorit y	Action	Close s	Effort	Success criterion
1	Produce a versioned specification for the coordinate-to-metric layer with a conformance test suite	G2	low	two independent implementations agree to within 1° on common pose files
2	Adopt the checklist of Table XI at submission	G2, G5	low	≥80% of new papers include all eight items
3	Produce an athletic benchmark with kinematics and phase ground truth, stratified by subgroup	G4	high	≥6 sports and one non-cyclical task; per-joint, per-plane agreement
4	Validate against motion capture using Bland-Altman limits and an ICC model	G5	high	bias and limits of agreement by joint, plane, and subgroup
5	Construct a causal lightweight estimator with calibrated depth uncertainty	G1	high	live tier (≥30 fps, ≤100 ms) on consumer CPU with usable depth intervals
6	Assess longitudinal baselines for each athlete	G6	medium	test-retest ICC ≥0.90 for the baseline deviation
7	Conduct a blinded practitioner decision-utility study	G5	high	observable decision difference compared to video review
8	Deploy on-device and audit subpopulation fairness	equity	medium	no unprocessed video leaves the device; subpopulation errors are reported

J. A Noted Limitation of Scope

Because the initial BlazePose description is available only as a preprint and has been excluded per our criteria, BlazePose will be described exclusively via peer-reviewed empirical studies. This is an intended effect of the preprint exclusion criterion, documented here. The same criterion excludes several recent efficiency-focused architectures that have a preprint as their sole public description, which may lead to an unintentional bias in the Stage I corpus towards publications from venues with formal proceedings. We accept this bias for the guarantee that all quantified statements in this survey come from refereed works.



X. CONCLUSION

This study examined 45 articles written between 2021 and 2026 in a four-stage pipeline: landmark detection, coordinate-to-metric transformation, sport-related use, and validation & implementation. There are six gaps present across all of these stages: the gap between accuracy and deployability (G1), lack of standardization of the translation phase (G2), focus on individual exercise (G3), disparity of data (G4), lack of independent validation (G5), and population rather than personalized range of reference (G6).

The key conclusion of the study is that accuracy—this most researched and optimized characteristic—is not the binding limitation on the process of adoption. The limitations are documentation, independent validation of kinematics, and specifying the capture parameters. The explanation for this lies in the error propagation analysis which shows that landmark error propagates into angle error through a configuration-dependent amplification factor.

The demonstration of the reference implementation shows the practical applicability of the proposed temporally valid tiers as well as the relevance of defining claims boundaries per se. The architectural framework for the research field is already developed. What remains to be done is a relatively small effort on the common benchmarks and reference implementations.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the faculty and the staff of the Department of Information Technology, Sandip Institute of Technology and Research Centre, Nashik. We thank the anonymous reviewers for their comments which have helped to improve the arguments of Section VI-E.

REFERENCES

- [1] A. Badiola-Bengoa and A. Mendez-Zorrilla, "A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise," *Sensors*, vol. 21, no. 18, p. 5996, 2021, doi: 10.3390/s21185996.
- [2] C. Aulton, L. Wakili, B. W. Strafford, K. Davids, and C.-Y. Chiu, "The application of deep learning human pose estimation in sport: A systematic review," *Sports Medicine—Open*, vol. 11, no. 1, 2025, doi: 10.1186/s40798-025-00953-3.
- [3] M. H. B. Uçar, S. Solak, A. O. Jimale, H. Ünal, and S. Eken, "A comprehensive survey on pose estimation and tracking in sports: Methodologies, datasets, challenges, and future directions," *Artificial Intelligence Review*, vol. 59, no. 2, 2025, doi: 10.1007/s10462-025-11344-1.
- [4] S. Edriss, C. Romagnoli, L. Caprioli, V. Bonaiuto, E. Padua, and G. Annino, "Commercial vision sensors and AI-based pose estimation frameworks for markerless motion analysis in sports and exercises: A mini review," *Frontiers in Physiology*, vol. 16, 2025, doi: 10.3389/fphys.2025.1649330.
- [5] C. Zheng, W. Wu, T. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2024, doi: 10.1145/3603618.
- [6] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–41, 2023, doi: 10.1145/3524497.
- [7] Y. Xu, Z. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 38 571–38 584, doi: 10.52202/068431-2795.
- [8] C. Zheng, S. Zhu, M. Mendieta, T. Yang, T. Chen, and Z. Ding, "3D human pose estimation with spatial and temporal transformers," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 11 636–11 645, doi: 10.1109/ICCV48922.2021.01145.
- [9] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 222–13 232, doi: 10.1109/CVPR52688.2022.01288.



- [10] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, 2021, doi: 10.1109/TPAMI.2019.2929257.
- [11] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7157–7173, 2023, doi: 10.1109/TPAMI.2022.3222784.
- [12] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. European Conf. Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, 2016, pp. 483–499, doi: 10.1007/978-3-319-46484-8_29.
- [13] K. Sun, B. Xiao, D. Liu, and Y. Wei, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5686–5696, doi: 10.1109/CVPR.2019.00584.
- [14] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "VNect: Real-time 3D human pose estimation with a single RGB camera," *ACM Trans. Graphics*, vol. 36, no. 4, pp. 1–14, 2017, doi: 10.1145/3072959.3073596.
- [15] H. Zhang, J. Cai, H. Qiu, Q. Lei, J. Liu, and M. Zhang, "A survey of deep learning-based action quality assessment: Methods and applications," *Journal of Big Data*, vol. 13, no. 1, 2026, doi: 10.1186/s40537-026-01409-5.
- [16] S. Mroz, N. Baddour, C. McGuirk, P. Juneau, A. Tu, K. Cheung, and E. Lemaire, "Comparing the quality of human pose estimation with BlazePose or OpenPose," in *Proc. 4th Int. Conf. Bio-Engineering for Smart Technologies (BioSMART)*, 2021, pp. 1–4, doi: 10.1109/BioSMART54244.2021.9677850.
- [17] M. Mundt, Z. Born, M. Goldacre, and J. Alderson, "Estimating ground reaction forces from two-dimensional pose data: A biomechanics-based comparison of AlphaPose, BlazePose, and OpenPose," *Sensors*, vol. 23, no. 1, p. 78, 2022, doi: 10.3390/s23010078.
- [18] S. Dill, A. Ahmadi, M. Grimmer, D. Haufe, M. Rohr, Y. Zhao, M. Sharbafi, and C. Hoog Antink, "Accuracy evaluation of 3D pose reconstruction algorithms through stereo camera information fusion for physical exercises with MediaPipe pose," *Sensors*, vol. 24, no. 23, p. 7772, 2024, doi: 10.3390/s24237772.
- [19] F. Kahl, P. Wegner, M. Kapsecker, L. Nissen, J. Faber, S. M. Jonas, and L. Reimer, "A comparative assessment of accuracy in video-based monocular human pose estimation frameworks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 7, pp. 8470–8484, 2026, doi: 10.1109/TPAMI.2026.3672463.
- [20] S. Du, Z. Yuan, and T. Ikenaga, "Kinematics-aware spatial-temporal feature transform for 3D human pose estimation," *Pattern Recognition*, vol. 150, p. 110316, 2024, doi: 10.1016/j.patcog.2024.110316.
- [21] Y. Liu and Z. Zhang, "STGFormer: Spatio-temporal GraphFormer for 3D human pose estimation in video," *Pattern Recognition*, vol. 171, p. 112239, 2026, doi: 10.1016/j.patcog.2025.112239.
- [22] B. Lang and M. C. Chuah, "Event-guided video transformer for end-to-end 3D human pose estimation," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, 2025, pp. 5114–5124, doi: 10.1109/WACV61041.2025.00500.
- [23] 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, doi: 10.1016/j.neucom.2024.128049.
- [24] X. Yan, B. Liu, and G. Qu, "Efficient monocular human pose estimation based on deep learning methods: A survey," *IEEE Access*, vol. 12, pp. 72 650–72 661, 2024, doi: 10.1109/ACCESS.2024.3399222.
- [25] A. El Kaid and K. Baïna, "A systematic review of recent deep learning approaches for 3D human pose estimation," *Journal of Imaging*, vol. 9, no. 12, p. 275, 2023, doi: 10.3390/jimaging9120275.
- [26] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, and J. Luo, "Anatomy-aware 3D human pose estimation with bone-based pose decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 198–209, 2022, doi: 10.1109/TCSVT.2021.3057267.



- [27] X. Zhu, X. Zeng, and W. Ma, "Lightweight cross-fusion network on human pose estimation for edge device," *IEEE Access*, vol. 11, pp. 134 899–134 907, 2023, doi: 10.1109/ACCESS.2021.3065574.
- [28] X. Lou, X. Lin, H. Zeng, and X. Zhu, "LAR-Pose: Lightweight human pose estimation with adaptive regression loss," *Neurocomputing*, vol. 633, p. 129777, 2025, doi: 10.1016/j.neucom.2025.129777.
- [29] X. Chen, J. Zhang, K. Wang, P. Wei, and L. Lin, "Multi-person 3D pose estimation with occlusion reasoning," *IEEE Trans. Multimedia*, vol. 26, pp. 878–889, 2024, doi: 10.1109/TMM.2023.3272736.
- [30] J. Tu, G. Wu, and L. Wang, "Dual graph networks for pose estimation in crowded scenes," *Int. J. Computer Vision*, vol. 132, no. 3, pp. 633–653, 2024, doi: 10.1007/s11263-023-01901-y.
- [31] J. Choi, E. Ha, M. Son, J. Jeon, and J. Kim, "Human joint angle estimation using deep learning-based three-dimensional human pose estimation for application in a real environment," *Sensors*, vol. 24, no. 12, p. 3823, 2024, doi: 10.3390/s24123823.
- [32] L. Ciardiello, P. Agnello, M. Petyx, F. Martinelli, M. Cesarelli, A. Santone, and F. Mercaldo, "A method for human pose estimation and joint angle computation through deep learning," *Journal of Imaging*, vol. 12, no. 4, p. 157, 2026, doi: 10.3390/jimaging12040157.
- [33] J. Stenum, C. Rossi, and R. T. Roemmich, "Two-dimensional video-based analysis of human gait using pose estimation," *PLOS Computational Biology*, vol. 17, no. 4, p. e1008935, 2021, doi: 10.1371/journal.pcbi.1008935.
- [34] J. Stenum, M. M. Hsu, A. Y. Pantelyat, and R. T. Roemmich, "Clinical gait analysis using video-based pose estimation: Multiple perspectives, clinical populations, and measuring change," *PLOS Digital Health*, vol. 3, no. 3, p. e0000467, 2024, doi: 10.1371/journal.pdig.0000467.
- [35] G. Rigoni, M. D. Vacche, C. Impoco, F. Spolaor, and Z. Sawacha, "Monocular gait analysis via 3D mesh reconstruction: Harmonizing body pose estimation with classical gait analysis frameworks," *Gait & Posture*, vol. 130, p. 110513, 2026, doi: 10.1016/j.gaitpost.2026.110513.
- [36] C. Li, K. Yick, P. Li, S. McErlain-Naylor, and J. Yip, "Validity of wearable inertial measurement units for monitoring lower-limb running biomechanics: A comparison with optical motion capture," *BMC Sports Science, Medicine and Rehabilitation*, vol. 18, 2026, doi: 10.1186/s13102-026-02091-z.
- [37] L. Wade, L. Needham, P. McGuigan, and J. Bilzon, "Applications and limitations of current markerless motion capture methods for clinical gait biomechanics," *PeerJ*, vol. 10, p. e12995, 2022, doi: 10.7717/peerj.12995.
- [38] L. D. Haberkamp, M. C. Garcia, and D. M. Bazett-Jones, "Validity of an artificial intelligence, human pose estimation model for measuring single-leg squat kinematics," *Journal of Biomechanics*, vol. 144, p. 111333, 2022, doi: 10.1016/j.jbiomech.2022.111333.
- [39] T. Baumgartner and S. Klatt, "Monocular 3D human pose estimation for sports broadcasts using partial sports field registration," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 5109–5118, doi: 10.1109/CVPRW59228.2023.00539.
- [40] T. Baumgartner, B. Paassen, and S. Klatt, "Extracting spatial knowledge from track and field broadcasts for monocular 3D human pose estimation," *Scientific Reports*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-41142-0.
- [41] L. M. Ølstad, E. E. Semb, S. Hjortland, and M. Steinert, "Computer vision 360 video analysis in sports: 3D athlete pose and rig motion estimation in olympic sailing," *Applied Sciences*, vol. 16, no. 3, p. 1386, 2026, doi: 10.3390/app16031386.
- [42] I. Coifman, M. Hakim, G. Weiss, and Y. Riemer, "Human pose estimation for automated biomechanical swimming analysis," *Sports Biomechanics*, pp. 1–24, 2026, doi: 10.1080/14763141.2026.2715647.
- [43] P. S. Glazier, "Markerless motion capture in sport: Panacea or pandora's box?" *Sports Biomechanics*, vol. 25, no. 2, pp. 165–172, 2026, doi: 10.1080/14763141.2025.2458485.
- [44] J. Aleksić, D. Kanevsky, D. Mesaroš, O. M. Knezevic, D. Cabarkapa, B. Bozovic, and D. M. Mirkov, "Validation of automated countermovement vertical jump analysis: Markerless pose estimation vs. 3D marker-based motion capture system," *Sensors*, vol. 24, no. 20, p. 6624, 2024, doi: 10.3390/s24206624.



- [45] K. Ksoll, R. Krätschmer, and F. Stöcker, "Validity of a convolutional neural network-based, markerless pose estimation system compared to a marker-based 3D motion analysis system for gait assessment—a pilot study," *Sensors*, vol. 25, no. 21, p. 6551, 2025, doi: 10.3390/s25216551.
- [46] S. D. Uhlich, A. Falisse, Ł. Kidziński, J. Muccini, M. Ko, A. S. Chaudhari, J. L. Hicks, and S. L. Delp, "OpenCap: Human movement dynamics from smartphone videos," *PLOS Computational Biology*, vol. 19, no. 10, p. e1011462, 2023, doi: 10.1371/journal.pcbi.1011462.
- [47] E. Y. Miller, T. Tan, A. Falisse, and S. D. Uhlich, "Integrating machine learning with musculoskeletal simulation improves OpenCap video-based dynamics estimation," *IEEE Trans. Biomed. Eng.*, pp. 1–10, 2026, doi: 10.1109/TBME.2026.3691975.
- [48] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986, doi: 10.1016/S0140-6736(86)90837-8.
- [49] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.
- [50] C. K. Ingwersen, C. M. Mikkelsen, J. N. Jensen, M. R. Hannemose, and A. B. Dahl, "SportsPose—a dynamic 3D sports pose dataset," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 5219–5228, doi: 10.1109/CVPRW59228.2023.00550.
- [51] C. Yeung, T. Suzuki, R. Tanaka, Z. Yin, and K. Fujii, "AthletePose3D: A benchmark dataset for 3D human pose estimation and kinematic validation in athletic movements," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025, pp. 5935–5946, doi: 10.1109/CVPRW67362.2025.00592.
- [52] T. Suzuki, R. Tanaka, C. Yeung, and K. Fujii, "AthleticsPose: Authentic sports motion dataset on athletic field and evaluation of monocular 3D pose estimation ability," in *Proc. 8th Int. ACM Workshop on Multimedia Content Analysis in Sports*, 2025, pp. 8–17, doi: 10.1145/3728423.3759402.

