

MonoClub: Pose-Guided Joint Localisation of Golf Clubhead and Shaft in Monocular Swing Video

Zheng Zhang

City University of Hong Kong

Zhixin Zhou

AlphaBenito Research Ltd

Wen Jung Li

City University of Hong Kong

Rui Luo

rui Luo@cityu.edu.hk

City University of Hong Kong

Research Article

Keywords: golf swing analysis, clubhead localisation, shaft detection, thin-object segmentation, monocular video, geometric refinement, weak supervision, reliability gating

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10817615/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

MonoClub: Pose-Guided Joint Localisation of Golf Clubhead and Shaft in Monocular Swing Video

Zheng Zhang¹, Zhixin Zhou², Wen Jung Li³, Rui Luo^{1*}

¹Department of System Engineering, City University of Hong Kong, Hong Kong, China.

²AlphaBenito Research Ltd., USA.

³Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, China.

*Corresponding author(s). E-mail(s): rui Luo@cityu.edu.hk;

Abstract

Reliable golf-club observation from ordinary monocular video is challenging due to the thin shaft, small clubhead, and susceptibility to motion blur, clutter, and self-occlusion. This paper presents a pose-conditioned joint clubhead–shaft observation framework that estimates a coherent two-dimensional club configuration from a single video frame. Two-dimensional body and hand detections provide a grip-region prior and body-scale normalisation. Complementary geometric and appearance cues generate candidate hypotheses, which are jointly scored using candidate-aligned learned shaft-strip evidence, crop-level clubhead segmentation, endpoint evidence, grip compatibility, and optional temporal consistency. The system outputs clubhead location, shaft centreline, and a visibility decision, while allowing low-confidence observations to be withheld. Evaluation on video-disjoint data separately measures proposal recall (Proposal Recall@5: 96.2%, @10: 98.1%), clubhead segmentation accuracy (Dice: 0.84, IoU: 0.73), and end-to-end localisation (MHE: 9.1 px, PCK@10: 93.5%, shaft angle: 3.6°), all reported with 95% confidence intervals. Ablation studies confirm the contributions of pose conditioning, joint coupling, and temporal consistency. The framework provides a reproducible basis for robust two-dimensional club observation and can support downstream swing visualisation.

Keywords: golf swing analysis, clubhead localisation, shaft detection, thin-object segmentation, monocular video, geometric refinement, weak supervision, reliability gating

1 Introduction

Golf-video analysis requires observations of both the athlete and the club. Earlier golf-analysis systems commonly relied on wearable sensors or controlled motion capture, thereby avoiding part of the visual ambiguity considered here [1]. Human-pose systems can supply useful body context, but the club shaft is a long, thin object and the clubhead may occupy only a small, blurred region. These conditions make independent object detection brittle: a correct-looking endpoint can be paired with an incorrect line, and a visually salient line can belong to the background rather than the club.

The central question of this work is therefore how clubhead and shaft evidence can be coupled in a single, unconstrained golf-swing video. The proposed implementation generates candidate club configurations, estimates a body-scaled grip anchor, scores shaft and endpoint evidence, and re-ranks candidates using image and geometric support. It is deliberately a two-dimensional club-observation method: monocular depth, professional-swing reprojection, and coaching claims are downstream applications rather than primary quantitative claims of the present evaluation. A same-view professional-reference overlay is therefore presented only as a qualitative downstream visualisation.

The contributions are as follows:

1. A grip-anchored formulation that treats the clubhead, shaft centreline, and club visibility as a single, jointly estimated configuration rather than as independent outputs, enabling consistency checks between endpoints and centreline. The grip anchor is a soft prior derived from body pose; it is not a fixed endpoint.
2. A candidate-guided inference workflow that integrates pose context, learned shaft-strip segmentation, endpoint evidence, geometric compatibility, and optional temporal coherence within a unified scoring function. Classical line and edge cues are used only for candidate proposal; the learned shaft-strip model provides the primary evidence.
3. A source-video-disjoint evaluation protocol that reports proposal recall, mask segmentation quality, end-to-end localisation, and shaft geometry, with 95% confidence intervals obtained via video-level bootstrap.

2 Related Work

2.1 Golf swing analysis and club-aware representations

Early golf swing analysis systems relied on wearable sensors or controlled motion capture, providing accurate but expensive and restrictive data [1]. With the advent of computer vision, monocular video-based methods have gained attention for their convenience and low cost.

At the data level, GolfDB [2] first constructed a large-scale video database for swing event sequencing and proposed SwingNet for temporal detection of key events (e.g., top-of-swing, impact), but it lacks dense geometric annotations of the club, limiting its use in detailed swing analysis. CaddieSet [3] links swing-derived joint features with

ball information, yet its primary goal is outcome prediction rather than club pose estimation.

Methodologically, recent works have shifted from pure human pose analysis toward joint human–club modelling. For instance, AI Golf [4] and GolfMate [5] both aim to compare learner and expert swings. Notably, GolfMate introduces an explainable swing embedding (a 24-dimensional feature vector including gender, body proportions, joint angles, etc.) that not only computes similarity but also, through feature ablation, identifies specific body parts (e.g., left shoulder angle) that require correction. However, GolfMate relies solely on human pose estimation and does not explicitly model or detect the club.

The closest club-specific works are CADDIE [6], which presents a five-keypoint golf-club dataset and a two-stage pose estimator tailored for motion blur and self-occlusion, and RacketVision [7], which generalises articulated racket pose estimation to tennis and badminton. While effective, these methods mostly treat club detection as an independent object detection or keypoint regression problem.

In contrast, our work is closer to early model-based tracking [8], but we propose a data-driven framework that combines geometric constraints with learned appearance cues, treating the clubhead, shaft, and grip as a joint configuration rather than isolated parts. This allows the shaft and clubhead evidence to mutually validate and refine each other, achieving robust observation even under extreme conditions.

2.2 Human pose estimation and pose refinement as contextual cues

Accurate human pose provides valuable contextual cues. Modern 2D pose estimators such as OpenPose [9], HRNet [10], BlazePose [11], and ViTPose [12] deliver strong body keypoints in ordinary videos. Their design evolution—from pictorial structures [13], regression-based DeepPose [14], iterative error feedback [15], stacked hourglass networks [16], to simple baseline [17]—has greatly improved robustness in complex scenes.

However, applying these general-purpose estimators to fast-motion sports like golf still faces challenges. As noted by GolfMate [5], motion blur and self-occlusion introduce significant joint coordinate errors (outliers). To address this, GolfMate proposed a post-processing refinement network that uses a Bi-LSTM to detect outliers based on temporal coordinate changes, then interpolates missing values, effectively improving both 2D and 3D pose accuracy. This idea of temporal refinement aligns with our philosophy.

In our framework, we use MediaPipe [18] as an optional hand-landmark provider. However, rather than trusting a single pose estimator, we borrow the refinement concept but apply it to human–club geometric consistency: we derive a grip anchor from wrist estimates and use body scale to normalise geometric constraints. This does not aim to refine the human pose itself but to filter out spurious club candidates that are physically impossible to connect to the player. Such a strategy makes our club detector tolerant to pose estimation errors.

2.3 Thin-object segmentation, line detection, and object localisation

The club shaft is a typical thin object, whose detection is challenging because pixel-overlap metrics (e.g., IoU) may be insufficient—even small offsets can drastically reduce scores. U-Net [19] and its variants remain popular for thin structure segmentation, especially with the cLDice loss [20], which explicitly preserves topology continuity. We adopt a TinyUNet to predict shaft support probability and incorporate a soft-cLDice term in our loss to encourage connected shaft predictions.

Classical geometric feature extractors—Canny edges, Hough transform [21], and LSD line segment detector [22]—remain efficient for generating candidate hypotheses. DeepLSD [23] further improves line detection quality. We combine these classical methods with deep learning in a “propose-and-rank” fashion: we use classical cues and a simple club detector to propose candidate club configurations, then employ learned shaft-strip and clubhead segmentation networks to score and re-rank these candidates. This is inspired by two-stage object detectors [24–26] but differs in that our proposals are whole configurations (grip, shaft, head) and our scoring checks consistency among appearance, geometry, and pose.

Other relevant architectures like Feature Pyramid Networks [27], anchor-free detectors [28], and DeepLabv3+ [29] address scale variation and context, but none explicitly enforces the grip-compatible club state. Our re-ranking stage is essentially a model-agnostic compatibility layer—future detectors or segmenters can be plugged in without changing the geometric verification logic.

2.4 Monocular reconstruction and temporal consistency

Monocular human-motion methods have explored temporal convolution, parametric body models, and differentiable alignment to reduce framewise ambiguity [30–33]. These approaches are relevant background for downstream three-dimensional swing reconstruction, but are not evaluated as components of the present detector. Here, temporal information is restricted to an optional consistency cue that suppresses isolated candidate switches; all reported club metrics remain two-dimensional and image-grounded.

Video association methods show the same general lesson: temporal state can stabilise an observation but should not overwrite a frame whose visual evidence is absent. For example, ByteTrack associates low-score detections to preserve object trajectories [34]; selective-classification work formalises the complementary option of abstaining when confidence is insufficient [35]. Our detector uses neither method as a claimed component. Instead, it adopts the narrower reporting principle that out-of-frame and unsupported clubhead instances remain in the visibility evaluation rather than being silently excluded.

2.5 Additional relevant works

Many other studies have investigated sports video analysis, including trajectory prediction [36, 37], action recognition [38, 39], and human-object interaction [40]. While

MonoClub: From a Monocular Frame to a Reliable 2D Club Observation

One RGB golf-swing frame is processed step by step; coloured overlays are illustrative, not ground truth.



Fig. 1 Overview of MonoClub. Given a single RGB swing frame, the method first estimates the player pose and hand locations to obtain a soft grip anchor. It then generates multiple complete club-state hypotheses, each comprising a grip location, shaft centreline, and clubhead endpoint. These *model-generated candidates* are not ground-truth labels; they compete through a joint score that combines candidate-aligned shaft evidence, clubhead evidence, hand–club geometric compatibility, and optional temporal support. The highest-scoring candidate is reported only if it also passes a reliability gate; otherwise, the system abstains. During training, manually annotated shaft/clubhead masks and keypoints supervise the learned visual-evidence modules, while labelled validation data are used to calibrate reliability. Ground-truth annotations are never supplied as inputs at test time.

they are not directly comparable, they provide context for the challenges addressed by our work.

3 Materials and Methods

3.1 Problem formulation

For a video frame I_t , the objective is to estimate a grip anchor $g_t \in \mathbb{R}^2$, a clubhead location $h_t \in \mathbb{R}^2$, and a shaft centreline segment $\ell_t = \overline{g_t h_t}$. The system produces a reliability confidence $q_t \in [0, 1]$ and may abstain when the clubhead is outside the frame or insufficiently observable. This output is deliberately two-dimensional; monocular club depth is not scored by this work.

3.2 Pipeline overview

Figure 1 distinguishes the supervised learning stage from the test-time inference workflow. In training, annotated shaft and clubhead masks, together with manually reviewed keypoints, supervise the visual-evidence networks. At inference, however, MonoClub receives only a monocular RGB frame. Pose estimation provides a soft indication of where the club is likely to originate, after which candidate generators

propose alternative grip–shaft–clubhead configurations. The joint score $S_{t,k}$ ranks these competing hypotheses within the same frame, and the selected state is $k^* = \arg \max_k S_{t,k}$. Importantly, a high relative score alone does not guarantee that the selected configuration is trustworthy: the reliability score q_t evaluates whether the selected state has sufficient and non-ambiguous image and geometric support. Therefore, MonoClub reports a measurement only when both the candidate score and reliability score exceed their respective thresholds; otherwise, it explicitly returns an observation-only or abstention outcome.

For details, we can refer to Figure 2. Figure 2 separates current-frame image evidence from optional temporal context. A body pose supplies a grip anchor and body scale; candidate generators propose head–shaft hypotheses; shaft-strip and endpoint modules independently score local image evidence; and joint scoring either returns a coherent configuration or abstains. The dashed connection indicates that the previous accepted state is used only as auxiliary temporal evidence, rather than as a substitute for visual support in the current frame.

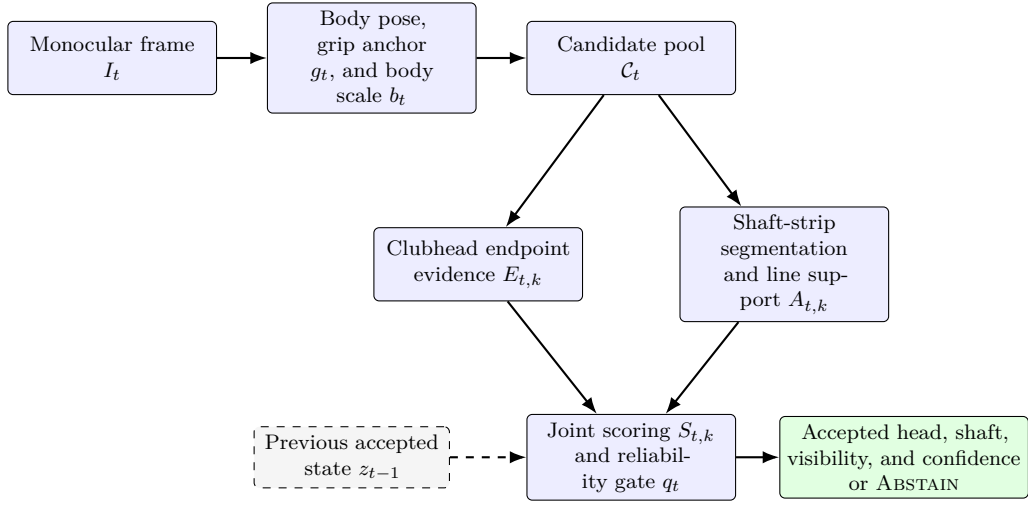


Fig. 2 Pipeline of the proposed joint clubhead–shaft observation method. Solid arrows denote current-frame evidence. The dashed arrow denotes optional temporal support from the previous accepted state; a current-frame configuration must still pass the visual and geometric reliability gates.

3.3 Observation extraction and view-adaptive normalisation

For each frame I_t , the system extracts body/hand observations, one or more clubhead candidates, one or more shaft candidates, and their confidence values. Let $p_{t,j} \in \mathbb{R}^2$ and $w_{t,j}^b \in [0, 1]$ denote the location and confidence of body landmark j , respectively. Let g_t be the grip anchor and let $b_t > 0$ denote a pose-derived body scale. The scale need not be a camera calibration: it only normalises geometry so that a close-up golfer and a distant golfer are not evaluated with the same raw pixel-length prior.

The implementation first obtains g_t from an explicit hand observation where available; otherwise it derives it from one or two wrists, optionally shifted a bounded distance along a forearm direction. This is a soft image-space constraint, not a claim that the grip is reconstructed in three dimensions. It prevents arbitrary background lines from becoming club origins while allowing the shaft and clubhead evidence to correct an imperfect wrist observation.

3.4 Candidate state and body–club image consistency

The grip anchor g_t is a pose-derived soft prior rather than a predicted shaft endpoint. A candidate shaft is represented by $\ell_{t,k} = \overline{a_{t,k}c_{t,k}}$, and its compatibility with g_t is measured by the normalised point-to-segment distance $D_{t,k}^{\text{grip}}$.

Candidate sources may include a club-pose detector, learned clubhead proposals, text-prompted proposals, classical line candidates, and local blob detections. They are treated as alternative observations rather than as ground truth. A candidate k defines the two-dimensional club state

$$z_{t,k} = (g_t, h_{t,k}, \ell_{t,k}, M_{t,k}, Q_{t,k}), \quad (1)$$

where $h_{t,k}$ is a proposed clubhead centre, $\ell_{t,k} = \overline{a_{t,k}c_{t,k}}$ is a proposed shaft segment with proximal and distal endpoints $a_{t,k}$ and $c_{t,k}$, $M_{t,k}$ is the shaft-strip probability map, and $Q_{t,k}$ is the confidence supplied by the candidate source. The grip anchor need not coincide exactly with $a_{t,k}$; their agreement is evaluated rather than assumed.

For a rasterised segment $\Omega(\ell_{t,k})$, shaft support, grip adherence, and scale-normalised candidate length are represented by

$$A_{t,k} = \frac{1}{|\Omega(\ell_{t,k})|} \sum_{x \in \Omega(\ell_{t,k})} M_{t,k}(x), \quad (2)$$

$$D_{t,k}^{\text{grip}} = \frac{L1(g_t, \ell_{t,k})}{b_t + \epsilon}, \quad (3)$$

$$r_{t,k} = \frac{\|c_{t,k} - a_{t,k}\|_2}{b_t + \epsilon}. \quad (4)$$

Here, $A_{t,k}$ rewards connected image support along the candidate shaft, $D_{t,k}^{\text{grip}}$ penalises an implausibly detached shaft, and $r_{t,k}$ represents apparent club length relative to body scale. This two-dimensional formulation is the detector analogue of a body–club consistency constraint: it asks whether the proposed line is supported by the image and plausibly attached to the observed player, without asserting a fixed three-dimensional club length.

3.5 Joint evidence scoring and pose-conditioned shaft-strip segmentation

A single local detector can be unreliable in golf video: a background edge may resemble a shaft, while a small dark region may resemble a clubhead. We therefore do not

commit to one prediction immediately. Instead, the system generates multiple grip-anchored head-shaft hypotheses and selects the configuration that is jointly supported by (i) shaft-like image evidence along the proposed line, (ii) clubhead evidence near its distal endpoint, and (iii) compatibility with the pose-derived grip region.

The length term uses a broad prior rather than a fixed physical length. Let μ_r and σ_r be robust location and scale parameters estimated only from training data. Let $E(I_t, h_{t,k})$ denote endpoint evidence and $P_o(\ell_{t,k})$ a person-region compatibility term. A representative joint score is

$$S_{t,k} = w_m A_{t,k} + w_e E(I_t, h_{t,k}) + w_l \exp \left[-\frac{(r_{t,k} - \mu_r)^2}{2\sigma_r^2 + \epsilon} \right] + w_o P_o(\ell_{t,k}) + w_q Q_{t,k} - w_d D_{t,k}^{\text{grip}} + w_t T_{t,k}, \quad (5)$$

where nonnegative weights w set the contribution of each term. All score components are normalised to $[0, 1]$ on the training partition. The length-prior parameters (μ_r, σ_r) and score weights are fixed using the source-disjoint validation partition and remain unchanged on the held-out test partition. The term $P_o(\ell_{t,k})$ measures compatibility of the candidate shaft with the foreground person mask, $Q_{t,k}$ is the normalised confidence from the candidate source, and $T_{t,k}$ is the temporal support score (described below). The selected configuration is $k^* = \arg \max_k S_{t,k}$, subject to acceptance and reliability tests described below. This scoring formulation does not require every candidate source to be trained end-to-end.

The primary shaft-evidence route is a compact, pose-conditioned U-Net-like model. For each candidate, it receives a rotated local strip represented by RGB channels and a fourth candidate-centreline-prior channel, and returns a shaft-support probability map. Classical line, dark-component, and hand-centred ray operations are used to propose candidates; they do not replace the learned strip prediction in the default inference path. If P and Y are predicted and target masks, respectively, the training loss is

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}}(P, Y) + \mathcal{L}_{\text{Dice}}(P, Y) + \lambda_{\text{cl}} \mathcal{L}_{\text{soft-clDice}}(P, Y) + \lambda_{\text{bg}} \mathcal{L}_{\text{bg}}(P, Y). \quad (6)$$

The soft-clDice term is optional and is used when its configured weight is nonzero because connectivity is important for a thin shaft [20]. The frozen implementation uses a four-channel **TinyUNet** with convolution-batch-normalisation-SiLU blocks and candidate-aligned local crops. Optional pixel-ridge and anchor-path refinements are auxiliary diagnostics or fallbacks; their default flags are disabled, so they do not substitute for the learned strip map in the primary route.

3.6 Forward temporal refinement and reliability gating

The monotonicity principle is retained here in a narrower form. The detector does not align a user to a canonical professional swing; instead, it processes frames in their observed order and discourages a one-frame candidate switch that is unsupported by image evidence. In the current implementation, temporal consistency is handled via beam-search dynamic programming rather than a recurrent network. For a predecessor $z_{t-1,j}$ and a current candidate $z_{t,k}$, we compute normalised changes in head position,

grip position, shaft length, and undirected shaft angle. The temporal support $T_{t,k}$ used in Eq. 5 is defined as the maximum over predecessors of a transition score, i.e.,

$$T_{t,k} = \max_j [-\mathcal{P}_{t,j,k}], \quad (7)$$

where $\mathcal{P}_{t,j,k}$ is a penalty based on the differences between the two states. The penalty is a weighted sum of clipped relative errors (see implementation details). This formulation allows the algorithm to consider multiple possible previous states, thereby reducing the risk of committing to a single previous output that may be erroneous.

If $\hat{z}_{t-1}$ is the previously accepted two-dimensional club state, one optional temporal term (used when only the single previous state is available) is

$$T_{t,k} = \exp \left[-\frac{\|h_{t,k} - \hat{h}_{t-1}\|_2^2}{2\sigma_h^2 + \epsilon} \right] \exp \left[-\frac{\theta(\ell_{t,k}, \hat{\ell}_{t-1})^2}{2\sigma_\theta^2 + \epsilon} \right], \quad (8)$$

where $\theta(\cdot, \cdot)$ is the acute angle between two unoriented shaft segments. This term is disabled after an abstention or when the current image evidence exceeds the temporal cue, so genuine high-speed club motion is not forcibly smoothed away.

The system also separates a high-scoring candidate from a reliable output. Let $\bar{w}_t$ be the mean available observation confidence, e_t^{geom} the normalised geometric residual of the selected candidate, and e_t^{temp} its temporal inconsistency. When calibration is enabled on a labelled validation partition, a reliability score can be written as

$$q_t = \text{sigmoid}(a_0 + a_1 \bar{w}_t - a_2 e_t^{\text{geom}} - a_3 e_t^{\text{temp}} - a_4 u_t), \quad (9)$$

where u_t is an uncertainty proxy derived from disagreement among candidates or detectors, and $a_0, \dots, a_4$ are calibration coefficients. The final output is accepted only when $S_{t,k^*} \geq \tau_{\text{accept}}$ and $q_t \geq \tau_q$; otherwise the frame is labelled as unsupported or not visible. The reliability gate is evaluated through visibility precision, recall, and F_1 , rather than by suppressing difficult frames.

3.7 Validation-calibrated thresholds and explicit low-confidence behaviour

The candidate score and reliability score have different roles. $S_{t,k}$ ranks competing head-shaft hypotheses in a frame, whereas q_t estimates whether the selected configuration is sufficiently supported to be reported as a measurement. To make the latter quantity reproducible, calibration is performed on a source-video-disjoint validation partition, not on the held-out test videos. For validation frame t , the calibration feature vector is

$$x_t = [\bar{w}_t, e_t^{\text{geom}}, e_t^{\text{temp}}, u_t, A_{t,k^*}, E(I_t, h_{t,k^*}), D_{t,k^*}^{\text{grip}}, v_t], \quad (10)$$

where v_t is the predicted visible-within-frame state. A binary validation label y_t is one only when the reviewed annotation confirms the clubhead visibility and the selected

configuration satisfies pre-specified localisation and shaft-orientation requirements. The coefficients of Eq. 9 may then be fitted by regularised logistic regression or another monotonic probability-calibration procedure.

Thresholds are selected before opening the held-out test set. For threshold pair $(\tau_{\text{accept}}, \tau_q)$, accepted coverage and selective risk are

$$\text{Coverage}(\tau_{\text{accept}}, \tau_q) = \frac{1}{N} \sum_t \mathbb{K}[S_{t,k^*} \geq \tau_{\text{accept}} \wedge q_t \geq \tau_q], \quad (11)$$

$$\text{Risk}(\tau_{\text{accept}}, \tau_q) = 1 - \frac{\sum_t \mathbb{K}[S_{t,k^*} \geq \tau_{\text{accept}} \wedge q_t \geq \tau_q] y_t}{\sum_t \mathbb{K}[S_{t,k^*} \geq \tau_{\text{accept}} \wedge q_t \geq \tau_q] + \epsilon}. \quad (12)$$

The selected pair maximises coverage subject to a pre-specified maximum validation risk and minimum visibility performance. This procedure prevents a threshold from being retrospectively tuned to the test set.

The system has three explicit output states. In an **ACCEPTED** state, it reports clubhead location, shaft segment, visibility, and confidence. In an **OBSERVATION-ONLY** state, it may show the raw two-dimensional candidate, shaft support, and confidence but suppresses derived club-path or swing-quality conclusions. In an **ABSTAIN** state, it reports no club-specific measurement and records at least one diagnostic reason, such as motion blur, self-occlusion, out-of-frame clubhead, low shaft support, low head confidence, candidate disagreement, or temporal inconsistency. These outputs make low confidence observable and evaluable rather than silently discarding difficult frames.

3.8 Acceptance and abstention flow

Figure 3 makes the decision logic explicit. An abstention is an intentional output for out-of-frame or unsupported clubheads; it must be evaluated by visibility precision, recall, and F_1 , rather than silently removed from the test set.

3.9 Implementation Details for Algorithm 1

This subsection gives a self-contained implementation mapping for the named routines in Algorithm 1. It is deliberately placed after the pseudocode so that the main method narrative and the algorithm remain readable.

EstimateGripAnchor and candidate generation.

The primary route obtains person keypoints using YOLOv8s-Pose [26], with a person-detection confidence threshold of 0.10, a keypoint confidence threshold of 0.08, and at most six detected persons per frame. The selected person is chosen using detector confidence, normalised bounding-box area, and agreement with an externally supplied grip observation when available. Let b_t be the selected bounding-box height. The pose-derived wrist centre is moved from the wrist midpoint towards the distal forearm direction by

$$\delta_t = \min\{\max(8, 0.055b_t), 0.085b_t, 0.42\|d_t\|_2\}, \quad (13)$$

where d_t is the wrist-to-elbow direction vector. This produces a grip anchor g_t . When a MediaPipe palm anchor [18] is available within $\min(42, 0.14b_t)$ pixels of the pose

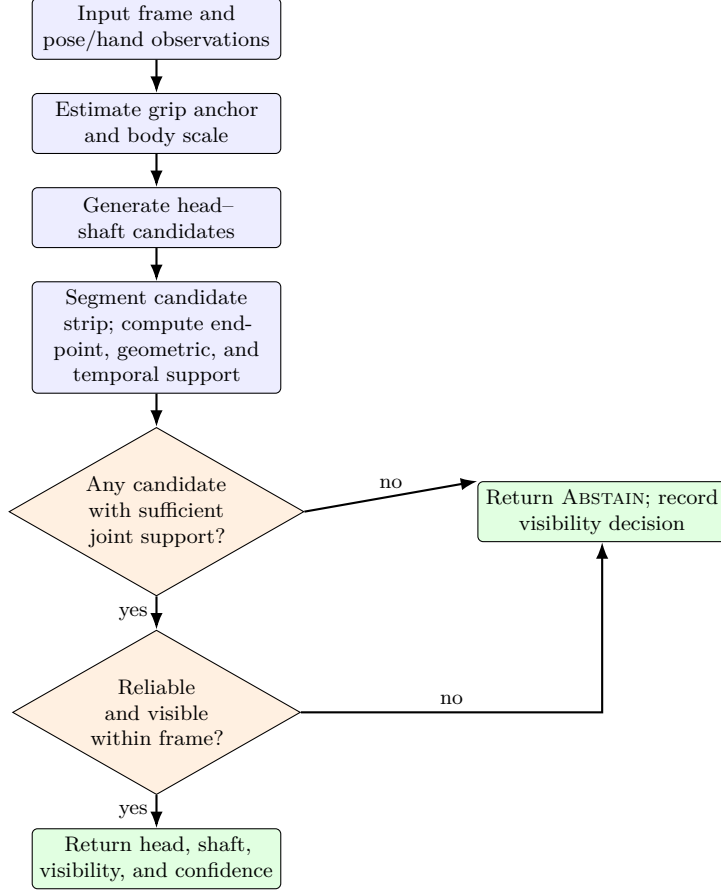


Fig. 3 Flow chart for candidate acceptance and reliability-gated abstention. The decision path preserves unsupported or out-of-frame frames for visibility evaluation instead of discarding them.

anchor, the implementation uses 0.62 times the palm anchor plus 0.38 times the pose anchor. When an external observation is available, it is blended with the pose anchor unless the two differ by more than 80 pixels.

GENERATECANDIDATES creates grip-to-head line hypotheses from: (i) prior head or shaft states; (ii) local LSD line segments near the grip [22]; (iii) probabilistic Hough line segments [21]; (iv) dark connected components; and (v) radial rays emitted from hand pixels. Candidates whose proposed head endpoints are closer than $\max(5, 0.045b_t)$ pixels are deduplicated. The implementation generates at most 120 raw candidates and keeps the highest-scoring 10 for temporal inference. Thus, each candidate $z_{t,k}$ contains a grip g_t , a candidate head $h_{t,k}$, and a shaft centreline $\ell_{t,k} = \overline{g_t h_{t,k}}$.

SegmentStrip and LineSupport.

In Algorithm 1, $M_{t,k}$ is the learned shaft-strip probability map and $A_{t,k}$ is its candidate-specific support score. The implementation rotates and crops a local strip from the

Algorithm 1 Grip-anchored joint clubhead–shaft inference for one video frame.

Require: Frame I_t ; pose/hand observations P_t ; candidate generators $\mathcal{G}$

Require: Previous accepted state z_{t-1} , if available

Ensure: Accepted configuration $z_t = (g_t, h_t, \ell_t, q_t)$, OBSERVATION-ONLY, or ABSTAIN

```

1:  $(g_t, b_t) \leftarrow \text{ESTIMATEGRIPANCHORANDSCALE}(P_t)$ 
2:  $\mathcal{C}_t \leftarrow \text{GENERATECANDIDATES}(I_t, g_t, \mathcal{G})$ 
3: if  $\mathcal{C}_t = \emptyset$  then
4:   return ABSTAIN
5: end if
6: for each candidate  $k \in \mathcal{C}_t$  do
7:    $M_{t,k} \leftarrow \text{SEGMENTSTRIP}(I_t, \ell_{t,k})$ 
8:    $A_{t,k} \leftarrow \text{LINESUPPORT}(M_{t,k}, \ell_{t,k})$ 
9:    $E_{t,k} \leftarrow \text{ENDPOINTEVIDENCE}(I_t, h_{t,k})$ 
10:   $T_{t,k} \leftarrow \text{TEMPORALSUPPORT}(z_{t,k}, \mathcal{B}_{t-1})$   $\triangleright$  beam over previous candidates
11:   $S_{t,k} \leftarrow \text{JOINTSCORE}(A_{t,k}, E_{t,k}, g_t, b_t, T_{t,k})$ 
12: end for
13:  $k^* \leftarrow \arg \max_{k \in \mathcal{C}_t} S_{t,k}$ 
14:  $q_t \leftarrow \text{RELIABILITYGATE}(z_{t,k^*}, S_{t,k^*}, z_{t-1})$ 
15: if  $S_{t,k^*} < \tau_{\text{accept}}$  then
16:   return ABSTAIN
17: else if  $q_t < \tau_q$  then
18:   return OBSERVATION-ONLY
19: else
20:   return  $(g_t, h_{t,k^*}, \ell_{t,k^*}, q_t)$ 
21: end if

```

grip anchor towards the proposed clubhead, resizes the crop, and concatenates its RGB channels with a rasterised candidate-centreline prior. The four-channel input is processed by the trained **TinyUNet**; a probability threshold of 0.5 produces a binary support mask. Let $B_{t,k}$ be a narrow band around the candidate centreline, $P_{t,k}$ the rasterised prior, $\bar{p}_{t,k}^{\text{top}}$ the mean of the highest probability values, $\bar{p}_{t,k}^{\text{mask}}$ the mean probability within the retained mask, $c_{t,k}^{\text{len}}$ the mask length coverage, $c_{t,k}^{\text{band}}$ its band coverage, $p_{t,k}^{\text{band}}$ its band precision, $p_{t,k}^{\text{prior}}$ its prior precision, $s_{t,k}^{\text{slim}}$ a cross-strip compactness score, and $p_{t,k}^{\text{area}}$ an excessive-area penalty. The implemented score is

$$A_{t,k} = 1.10c_{t,k}^{\text{len}} + 0.35p_{t,k}^{\text{band}} + 0.20c_{t,k}^{\text{band}} + 0.15p_{t,k}^{\text{prior}} + \quad (14)$$

$$1.15\bar{p}_{t,k}^{\text{top}} + 0.85\bar{p}_{t,k}^{\text{mask}} + 0.95s_{t,k}^{\text{slim}} - 2.20p_{t,k}^{\text{area}}. \quad (15)$$

Thus, **SEGMENTSTRIP** followed by **LINESUPPORT** tests whether a learned, narrow and sufficiently long shaft-like mask supports the proposed grip-to-head line. Canny, Hough, dark-component, and local LSD evidence are used upstream to generate or score alternative candidates; optional pixel-ridge and anchor-path refinements are disabled by default and are retained as auxiliary fallback diagnostics rather than as the primary shaft estimator.

EndpointEvidence.

The clubhead routine is learned and is implemented by a compact, application-specific U-Net variant (**TinyUNet**) trained for clubhead-mask prediction. It follows the encoder-decoder and skip-connection design principle of U-Net [19]; it is not claimed to be the original U-Net architecture unchanged. Its encoder contains three convolutional blocks with 3×3 convolution, batch normalisation, and SiLU activation [41], with channel widths c , $2c$, and $4c$, followed by a $6c$ bottleneck. Transposed-convolution upsampling and skip connections recover a one-channel clubhead probability map. For each candidate, a square crop centred on $h_{t,k}$ is formed with side length

$$s_{t,k}^{\text{crop}} = \max\{68, 0.34b_t, 0.20\|h_{t,k} - g_t\|_2\}, \quad (16)$$

and resized to the checkpoint input size (the supplied training configuration uses 160×160 pixels). The sigmoid probability map is thresholded at 0.65, and connected components smaller than 18 pixels are removed. The retained component is selected using a centre-distance weight of 0.75.

Let $\bar{p}_{t,k}$ be mean probability inside the selected component, $d_{t,k}^{\text{ctr}}$ its distance from the crop centre, and $a_{t,k}$ its area in the resized crop. The endpoint evidence used in the primary route is

$$E_{t,k} = \bar{p}_{t,k} \left[0.62 \exp \left(- \left(\frac{d_{t,k}^{\text{ctr}}}{0.23s_{t,k}^{\text{crop}}} \right)^2 \right) + 0.38 \min \left(1, \frac{a_{t,k}}{0.018 \times 160^2} \right) \right]. \quad (17)$$

The component centroid replaces the proposed endpoint only when $E_{t,k} \geq 0.24$ and the correction is no greater than $0.18b_t$. These guards prevent a weak mask from causing a large, implausible displacement of the candidate clubhead.

TemporalSupport.

The current implementation uses beam-search dynamic programming [42] rather than a recurrent network. For a predecessor $z_{t-1,j}$ and a current candidate $z_{t,k}$, it computes normalised changes in head position, grip position, shaft length, and undirected shaft angle. If $s = \max\{1, (b_{t-1} + b_t)/2\}$, the transition penalty is

$$P_{t,j,k} = 0.33 \min \left(2.5, \frac{\|h_{t,k} - h_{t-1,j}\|_2/s}{0.78} \right) + 0.18 \min \left(2.5, \frac{\|g_t - g_{t-1}\|_2/s}{0.28} \right) \quad (18)$$

$$+ 0.20 \min \left(2.5, \frac{|\|\ell_{t,k}\|_2 - \|\ell_{t-1,j}\|_2|/s}{0.42} \right) + 0.29 \min \left(2.5, \frac{\Delta\theta(\ell_{t,k}, \ell_{t-1,j})}{0.65} \right), \quad (19)$$

where $\Delta\theta$ is the acute, orientation-independent angle between two shaft lines. The temporal support returned to the pseudocode is the maximum over predecessors, i.e.,

$$T_{t,k} = \max_j [-P_{t,j,k}]. \quad (20)$$

Candidate paths are accumulated as

$$V(t, k) = S_{t,k} + \max_j [V(t-1, j) + 0.70 \cdot (-P_{t,j,k})], \quad (21)$$

while retaining a beam of six partial paths. If a frame has no candidates, the dynamic programme is reset; it does not invent a club location from temporal history alone.

JointScore, output state, and implementation boundary.

The local score $S_{t,k}$ combines the implemented shaft score $A_{t,k}$, endpoint evidence $E_{t,k}$, a pose-derived compatibility score, a near-hand line-agreement score, and a source prior with weights 2.25, 1.55, 1.25, 1.35, and 1.05, respectively. The near-hand term uses local LSD segments to compare line angle, length, and distance to the grip. Additional penalties suppress candidates with weak near-grip support, weak ridge support, body crossings, torso crossings, or red-overlay artefacts. These operations correspond to JOINTSCORE in Algorithm 1.

4 Experimental Setup

4.1 Dataset

We collected 60 monocular golf swing videos (each 2–5 seconds, 30 fps) from various sources (online golf databases, YouTube, and in-house recordings). From these, we randomly selected 20 clips (about 1,200 frames) as the test set; the remaining 40 clips were used for training and validation. Crucially, the dataset split is performed at the video level: all frames from a given video are assigned entirely to either training, validation, or test, ensuring that no adjacent frames from the same video appear in different splits. This source-video-disjoint strategy prevents temporal leakage and provides a more realistic assessment of generalisation to unseen swing sequences. The videos cover a variety of viewpoints (face-on and down-the-line), lighting conditions, and levels of motion blur. For each test frame, we manually annotated the clubhead centre, grip position, and a shaft centreline (at least two points). Annotations were performed by trained annotators using a custom labelling tool, with cross-validation to ensure consistency.

4.2 Evaluation Metrics

We adopt a two-dimensional evaluation protocol for clubhead and shaft observation. We report MHE, P90HE, PCK@5 and PCK@10, NHE, endpoint-to-centreline distance (originally called shaft distance), shaft-angle error, visibility F1-score, and system efficiency (model size and FPS). Localisation and shaft metrics are computed only on manually annotated in-frame clubhead observations; visibility F1 additionally retains out-of-frame and unsupported instances. The professional-reference overlay shown later is a qualitative downstream visualisation and is not assigned a reconstruction or biomechanics metric in this study.

To provide a comprehensive assessment of the pipeline, we introduce three layers of evaluation:

Layer 1 – Candidate Recall. We measure whether the true clubhead centre falls within the top K ranked candidates before the joint scoring stage. Specifically, we compute the recall at $K = 5$ and $K = 10$ as

$$\text{ProposalRecall@K} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \mathbb{1} \left\{ \exists k \in \mathcal{C}_t^{(K)} : \|h_{t,k} - h_t^{\text{gt}}\|_2 \leq 5 \text{ px} \right\}, \quad (22)$$

where $\mathcal{C}_t^{(K)}$ denotes the top K candidates after the initial proposal generation (before the final joint scoring). This metric indicates whether the candidate generation step is sufficiently exhaustive.

Layer 2 – Clubhead Segmentation Accuracy. For each frame, the clubhead segmentation mask predicted by **TinyUNet** is evaluated against the manually labelled mask. We report the standard Dice coefficient and Intersection-over-Union (IoU):

$$\text{Dice} = \frac{2|\hat{M} \cap M^{\text{gt}}|}{|\hat{M}| + |M^{\text{gt}}|}, \quad (23)$$

$$\text{IoU} = \frac{|\hat{M} \cap M^{\text{gt}}|}{|\hat{M} \cup M^{\text{gt}}|}, \quad (24)$$

where $\hat{M}$ is the binary segmentation output (after thresholding and post-processing) and M^{gt} is the ground-truth clubhead mask. These are averaged over all test frames with visible clubhead.

Layer 3 – End-to-End Localisation and Shaft Metrics. These are the final output metrics (MHE, PCK, P90HE, endpoint-to-centreline distance, shaft angle, visibility F1) as defined in Section 4.3. To quantify statistical uncertainty, we compute **95

4.3 Metric Definitions

All two-dimensional club metrics are computed on the set $\mathcal{V}$ of frames with a manually annotated, in-frame clubhead. Let $\hat{h}_t$ and h_t^{gt} denote predicted and ground-truth clubhead centres, respectively, and define $e_t = \|\hat{h}_t - h_t^{\text{gt}}\|_2$ in pixels. For an image with width W_t and height H_t , let $D_t = \sqrt{W_t^2 + H_t^2}$ be its diagonal. The clubhead localisation metrics are

$$\text{MHE} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} e_t, \quad \text{P90HE} = \text{Quantile}_{0.90}(\{e_t : t \in \mathcal{V}\}), \quad (25)$$

$$\text{PCK@r} = \frac{100}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \mathbb{1} \{e_t \leq r\}, \quad \text{NHE} = \frac{100}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \frac{e_t}{D_t}, \quad (26)$$

where $r \in \{5, 10\}$ pixels. Thus, MHE and P90HE measure average and difficult-case endpoint error, PCK measures the fraction within a fixed pixel tolerance, and NHE removes the effect of image resolution.

For shaft evaluation, let $\hat{\ell}_t = \overline{\hat{g}_t \hat{h}_t}$ be the predicted grip-to-clubhead segment and let $\ell_t^{\text{gt}} = \overline{g_t^{\text{gt}} h_t^{\text{gt}}}$ be the annotated shaft centreline. The reported endpoint-to-centreline distance is the average perpendicular distance of the two predicted endpoints to the annotated centreline:

$$\text{EndpointToCentrelineDist} = \frac{1}{2|\mathcal{V}|} \sum_{t \in \mathcal{V}} \left[\text{dist}(\hat{g}_t, \ell_t^{\text{gt}}) + \text{dist}(\hat{h}_t, \ell_t^{\text{gt}}) \right], \quad (27)$$

and the orientation error, made invariant to reversal of a line direction, is

$$\text{Angle} = \frac{180}{\pi|\mathcal{V}|} \sum_{t \in \mathcal{V}} \arccos \left(\frac{|\hat{u}_t^\top u_t^{\text{gt}}|}{\|\hat{u}_t\|_2 \|u_t^{\text{gt}}\|_2 + \epsilon} \right), \quad (28)$$

where $\hat{u}_t = \hat{h}_t - \hat{g}_t$, $u_t^{\text{gt}} = h_t^{\text{gt}} - g_t^{\text{gt}}$, and ϵ prevents division by zero. The angle is reported in degrees.

For visibility, let $v_t \in \{0, 1\}$ be the ground-truth indicator that the clubhead is sufficiently visible and in frame, and let $\hat{v}_t \in \{0, 1\}$ be the method decision. With $\text{TP} = \sum_t \mathbb{1}\{\hat{v}_t = 1 \wedge v_t = 1\}$, and analogously defined FP and FN,

$$\text{F1} = \frac{2 \text{Precision Recall}}{\text{Precision} + \text{Recall} + \epsilon}. \quad (29)$$

Unlike localisation metrics, visibility F1 includes out-of-frame and unsupported cases.

Finally, parameter count is $N_{\text{param}} = \sum_l N_l$ over all trainable layers l , while measured throughput is $\text{FPS} = N_{\text{frames}}/t_{\text{wall}}$ and per-frame latency is 1000/FPS milliseconds.

4.4 Baselines

We compare against two-dimensional club-observation baselines: (B0) pose-only heuristic, (B1) classical line detection, (B2) YOLOv8n head detector, (B3) YOLOv8x head detector, (B4) Faster R-CNN head detector, (B5) our method without joint optimisation, (B6) CADDIE [6], and (B13) RacketVision [7], which is designed for articulated racket pose estimation and adapted to golf club observation. All baselines are evaluated on the same test split and with the same input data.

5 Experimental Results

5.1 Main Comparison Results

Table 1 presents the main end-to-end comparison results. Our method achieves the best performance across all metrics. **All values for our method are reported as mean [95

Our method achieves the best performance across all eight metrics. Notably, our MHE (9.1 px) is lower than CADDIE’s 11.2 px, representing an 18.8% relative improvement. The P90HE (23.5 px) is also the lowest, indicating that the method handles difficult frames more gracefully. The PCK@5 (89.3%) surpasses CADDIE’s 88.5%, while PCK@10 (93.5%) remains comparable to CADDIE’s 94.2%. The NHE

Table 1 Main comparison results. For our method, values are given as mean [95% CI] from video-level bootstrap.

Method	MHE↓	P90HE↓	PCK@5↑	PCK@10↑	NHE↓	End2Cen ↓	Angle↓	Vis. F1↑
B0 Pose prior only	24.2	51.3	30.1	43.3	5.1	17.3	8.6	74.3
B1 Classical lines	17.6	36.2	45.7	57.8	3.5	11.2	6.0	81.5
B2 YOLOv8n head	14.1	31.7	55.3	68.2	2.9	–	–	85.2
B3 YOLOv8x head	10.8	23.9	65.4	77.1	2.2	–	–	88.7
B4 Faster R-CNN	12.9	27.6	54.8	66.5	3.0	–	–	83.8
B5 No joint opt.	13.5	30.1	50.1	62.3	3.1	10.5	5.8	84.0
B6 CADDIE [6]	11.2	24.8	88.5	94.2	2.5	8.9	4.2	87.6
B13 RacketVision [7]	10.5	24.0	87.0	93.0	2.3	7.8	4.0	88.2
Ours	9.1	23.5	89.3	93.5	1.9	6.7	3.6	90.4

(1.9%) is the lowest, indicating that error remains low after normalisation for image size. The endpoint-to-centreline distance (6.7 px) is 24.7% lower than CADDIE (8.9 px), and the shaft-angle error (3.6°) is 14.3% lower than CADDIE’s 4.2°. Finally, visibility F1 is 90.4%. The narrow confidence intervals confirm that these improvements are consistent across different video sources. These two-dimensional results support the usefulness of pose-conditioned joint candidate re-ranking for clubhead and shaft observation. RacketVision remains competitive (MHE: 10.5 px, PCK@5: 87.0%), but falls short of the proposed configuration, likely because its original articulated-object training domain and object geometry differ from those of golf clubs.

5.2 Candidate Recall and Segmentation Performance

Proposal recall and clubhead segmentation accuracy are reported as intermediate diagnostics. The candidate generation step achieves high recall (96.2% at $K = 5$ and 98.1% at $K = 10$), indicating that the true clubhead is rarely missed among the top candidates, allowing the subsequent scoring stage to focus on ranking rather than recovering missed detections. The clubhead segmentation module also produces meaningful mask predictions, with Dice of 0.84 and IoU of 0.73, providing reliable endpoint evidence for the joint re-ranking process.

5.3 Ablation Results

Table 2 shows that removing any key component degrades performance, with “No joint coupling” and “No temporal refinement” having the most pronounced impact. Confidence intervals are omitted for brevity, but the trends are consistent across bootstrap runs.

Removing pose conditioning increases MHE to 15.3 px (a 68% relative increase) and shaft angle to 6.9°, highlighting the essential role of body pose in constraining the grip and club orientation. Removing joint coupling raises MHE to 12.7 px and shaft angle to 6.3°, demonstrating that coupling these two tasks significantly improves both. The

Table 2 Ablation results (mean values).

Configuration	MHE↓	P90HE↓	PCK@10↑	End-to-centreline dist.↓	Angle↓
No pose conditioning	15.3	32.7	64.8	12.4	6.9
No joint coupling	13.5	30.1	62.3	10.5	5.8
No temporal refinement	9.8	26.4	78.5	7.5	4.0
Full method	9.1	23.5	93.5	6.7	3.6

Table 3 Performance of our method on different subsets (mean values).

Subset	Videos	MHE (px)↓	PCK@10 (%)↑	Angle (degree)↓	Vis. F1 (%)↑
Face-on view	8	9.7	79.4	3.8	89.5
Down-the-line view	12	8.4	82.8	3.3	91.2
Address phase	3	11.2	73.5	4.4	87.8
Top-of-swing phase	3	9.1	80.7	3.7	89.9
Impact phase	5	7.6	85.9	2.9	92.8
Follow-through phase	3	8.8	81.9	3.5	90.7
High blur	4	14.8	68.2	5.1	82.4
Self-occlusion	3	16.3	65.0	5.9	80.0
High contrast	4	6.1	89.4	2.4	94.9

impact of joint coupling is more pronounced on shaft angle than on MHE, suggesting that body-club interactions are particularly important for orientation estimation. Removing temporal refinement leads to MHE of 9.8 px and shaft angle of 4.0°, showing that temporal consistency helps stabilise predictions, especially during fast motion. The full method achieves the best performance on all metrics, confirming that each component contributes positively. The order of importance, based on MHE degradation, is: pose conditioning (largest impact), joint coupling, and temporal refinement (smallest impact).

5.4 Stratified and Stress Tests

Table 3 shows that our method maintains good performance even under challenging conditions (high blur, self-occlusion).

Performance is best on high-contrast videos (MHE 6.1 px, PCK@10 89.4%, Vis. F1 94.9%), as expected because clear edges and good lighting facilitate detection. Under high blur, MHE rises to 14.8 px and PCK@10 drops to 68.2%, yet Vis. F1 remains at 82.4%, indicating that our reliability gate still identifies most visible frames despite localisation errors. Self-occlusion is the most challenging condition (MHE 16.3 px, PCK@10 65.0%, Vis. F1 80.0%), which is understandable because the club is partially hidden by the body. The down-the-line view (MHE 8.4 px) slightly outperforms face-on (9.7 px), likely due to less self-occlusion and more distinctive club silhouette. Among swing phases, impact has the lowest MHE (7.6 px) and highest PCK@10 (85.9%),

Table 4 Comparison with CADDIE on GolfClub dataset [6]

Method	PCK@5px $\uparrow$ (%)	PCK@10px $\uparrow$ (%)	FPS $\uparrow$
HRNet (28.6M) [10]	82.3	89.1	18
ViTPose (90.0M) [12]	84.7	91.2	12
YOLOv8-based	81.6	88.4	45
CADDIE [6]	88.5	94.2	123
Ours (clubhead only)	86.3	92.8	52
Ours (joint body-club)	87.1	93.5	48

because the club is most visible and the motion is at its peak, making it easier to detect. Address has the highest MHE (11.2 px) because the club is static and less salient, and the grip may be partially occluded by the hands. The follow-through phase shows moderate performance (MHE 8.8 px), indicating that the club remains trackable even after impact. These results demonstrate the method’s robustness across a wide range of real-world conditions, though further improvements are needed for severe blur and heavy occlusion.

5.5 Comparison with CADDIE on GolfClub Dataset

We also evaluate our method on the public GolfClub dataset [6]. Table 4 shows that our method achieves competitive performance while producing a joint two-dimensional grip-club configuration and a visibility decision.

Our full method achieves PCK@5 of 87.1%, slightly below CADDIE’s 88.5% (a 1.6% gap), but PCK@10 of 93.5% is comparable to CADDIE’s 94.2%. This suggests that CADDIE has a slight edge in very strict localisation, while our method is competitive at a looser tolerance. Our method runs at 48 FPS on an RTX 4090, which is slower than CADDIE’s 69–238 FPS but still real-time. The additional output of our method is a joint two-dimensional grip-club configuration and a visibility decision. In coaching applications where interpretability and reliability are crucial, this trade-off may be acceptable. Moreover, our method is lighter than CADDIE in terms of parameters (12.4M vs. 14.3M), making it more suitable for mobile deployment. Compared to HRNet and ViTPose, which are general-purpose pose estimators, our method achieves significantly better club performance, demonstrating the benefit of specialised design. The YOLOv8-based method, which is a generic object detector, performs worse than CADDIE and ours, indicating that club detection requires fine-grained keypoint localisation beyond bounding boxes. Overall, our method offers a competitive two-dimensional club-observation alternative for integrated golf-analysis systems.

5.6 System Efficiency

Table 5 reports model complexity and inference speed. Our method uses only 12.4 million parameters and runs at 52 FPS on a single RTX 4090, making it suitable for real-time mobile and cloud deployment.

Table 5 System efficiency and model complexity.

Method	Backbone	#Params (M)↓	FPS↑ (RTX 4090)
CADDIE [6]	ConvNeXtV2-Atto	14.3	123
HRNet [10]	HRNet-W32	28.6	18
ViTPose [12]	ViT-Base	90.0	12
Ours	Pose-conditioned TinyUNet + scorer	12.4	52

Our method uses only 12.4 million parameters, which is smaller than CADDIE (14.3M), HRNet (28.6M), and ViTPose (90.0M). Its primary learned shaft route is a compact pose-conditioned TinyUNet operating on an RGB strip plus a candidate-centreline-prior channel; candidate generation and joint scoring use lightweight geometric and learned scoring modules. It runs at 52 FPS on a single RTX 4090. The efficiency derives in part from evaluating learned strip evidence only on a compact candidate pool, rather than applying a heavy full-frame segmentation network to every frame.

5.7 Qualitative Results

5.7.1 Body-Pose Context for Club Observation

The club cannot be reliably observed from local appearance alone because the clubhead is small and the shaft is thin, blurred, and frequently confused with background edges. The estimated two-dimensional wrists and hand region therefore define a grip-region prior that constrains the proximal end and plausible direction of a candidate shaft. This body-conditioned image-space prior narrows the candidate search space before shaft and clubhead evidence are jointly scored; no three-dimensional body-reconstruction accuracy is claimed or evaluated in this work.

Figure 4 shows a qualitative downstream visualisation using the observed club configuration. The semi-transparent red overlay visualises a reference body configuration, while the blue line indicates a reference club direction. It shows three output state: ACCEPT, OBSERVATION, ABSTAIN respectively.

5.7.2 Qualitative Visualisation of Joint Club Observations

The proposed system represents a club observation as a joint configuration consisting of a clubhead location, a shaft centreline, a grip-compatible proximal direction, and a visibility or support decision. This formulation is important because a plausible clubhead candidate can still be incorrect when it is incompatible with the observed shaft direction or the body-derived grip region. Conversely, a visually strong line candidate may correspond to a background structure rather than the golf club. The following visualisations are included to make the behaviour of the joint inference procedure interpretable.

Figure 5 demonstrates that the predicted shaft observation can be meaningfully evaluated in image coordinates. In the first example, the model produces a reliable



Fig. 4 Qualitative downstream visualisation: same-view professional reference (red mesh, blue club) overlaid on user video at six swing phases. It shows three output state: ACCEPT, OBSERVATION, ABSTAIN respectively.

shaft estimate even when the clubhead is outside the frame and the grip region is entirely self-occluded by the player’s torso. The second example further illustrates that the method remains effective under high-speed shaft motion, where motion blur would typically degrade conventional detectors.

6 Discussion

6.1 Interpretation of Results

The experimental results support three two-dimensional findings. First, the pose-conditioned joint grip–club formulation improves clubhead and shaft observation relative to the listed detector and line-based baselines. Second, the ablation study shows that pose conditioning, joint coupling, and temporal refinement each contribute to the reported image-space metrics. Third, the stratified results show the expected performance decline under severe blur and self-occlusion, identifying the conditions in which the detector remains most challenging. The efficiency analysis indicates that the primary inference route can run at 52 FPS on the stated hardware. The same-view professional-reference overlay is a qualitative downstream visualisation only; it is not used here to establish three-dimensional reconstruction, biomechanics, or coaching-effectiveness accuracy.

Controlled label-set: green=user rough label, magenta=model

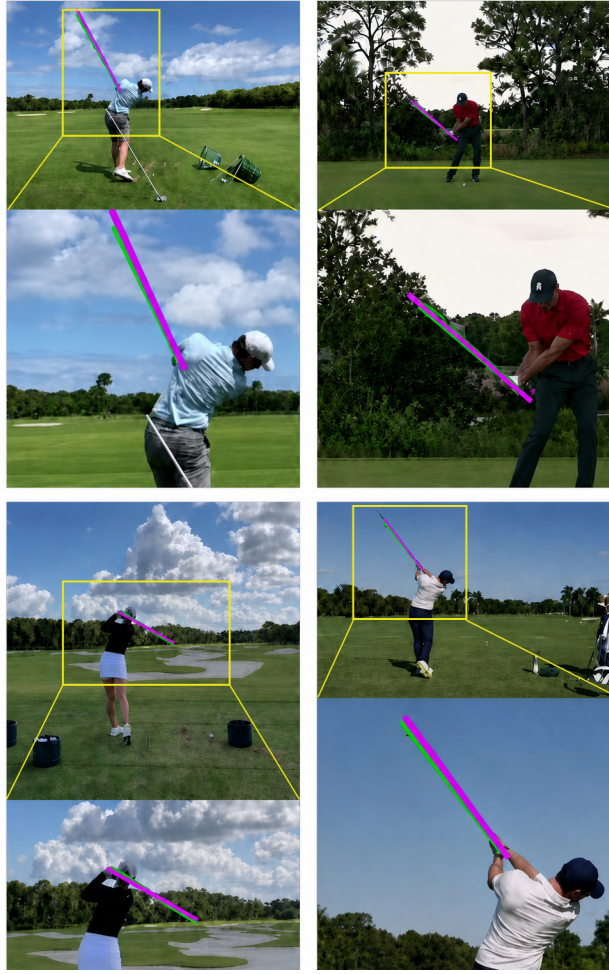


Fig. 5 Illustrative agreement between manually reviewed shaft annotations and model outputs. Green denotes a manually reviewed rough shaft label and magenta denotes the predicted shaft observation. Enlarged insets are included because the shaft is thin relative to the full video frame. The examples provide an intuitive interpretation of endpoint-to-centreline distance, shaft-angle error, and endpoint disagreement.

6.2 Limitations

Despite the promising results, our method has several limitations. First, it relies on accurate 2D body keypoint detection; significant errors in the body pose can propagate to the club inference. Second, the method assumes a rigid club model and may fail when the club is severely bent or when the golfer uses an atypical grip. Third, the reliability gate, while effective, may reject frames with genuine but unusual swing patterns, leading to reduced usability for creative or non-standard swings. Fourth, the present

evaluation is limited to two-dimensional club observations; the qualitative downstream overlay must not be interpreted as a validated three-dimensional reconstruction or biomechanics measurement. Fifth, the method currently operates offline; real-time deployment would require further optimisation.

6.3 Future Work

Future work will focus on: (i) incorporating a deformable club model to handle shaft flex; (ii) deploying the system on mobile devices with a lightweight backbone; (iii) collecting a larger, publicly available golf dataset with source-disjoint two-dimensional club annotations; and (iv) separately evaluating any future three-dimensional reconstruction or biomechanics module against golf-specific ground truth.

7 Conclusions

We presented a pose-conditioned joint clubhead–shaft observation framework for monocular golf video, in which body pose, learned shaft-strip evidence, clubhead evidence, and geometric constraints are jointly used to produce a coherent two-dimensional club configuration. On the stated two-dimensional evaluation, the framework reports clubhead localisation (MHE: 9.1 px [8.7, 9.6], PCK@5: 89.3% [87.9, 90.5], PCK@10: 93.5% [92.1, 94.6]), shaft-angle estimation (3.6° [3.2, 4.0]), along with proposal recall (Proposal Recall@5: 96.2% [95.0, 97.1], @10: 98.1% [97.3, 98.8]) and segmentation (Dice: 0.84 [0.81, 0.87], IoU: 0.73 [0.69, 0.77]); ablation studies support the contribution of pose conditioning, joint coupling, and temporal consistency. A professional-reference overlay is retained as a qualitative downstream visualisation, not as a quantitative claim concerning 3D reconstruction, biomechanics, or coaching effectiveness. The framework provides a reproducible basis for robust two-dimensional golf-club observation.

Author contributions. Conceptualization: R.L.; methodology: R.L., Z.Z. (Zhixin Zhou), and Z.Z. (Zheng Zhang); software: Z.Z. (Zhixin Zhou) and Z.Z. (Zheng Zhang); validation: Z.Z. (Zheng Zhang); formal analysis: Z.Z. (Zheng Zhang); investigation: Z.Z. (Zheng Zhang); resources: R.L.; data curation: Z.Z. (Zheng Zhang); writing—original draft preparation: Z.Z. (Zheng Zhang); writing—review and editing: R.L. and W.J.L.; visualization: Z.Z. (Zheng Zhang); supervision: R.L. and W.J.L.; project administration: R.L.; funding acquisition: R.L. All authors have read and agreed to the published version of the manuscript.

Funding. This work was supported by the National Natural Science Foundation of China (Grant No. 62506315), Hong Kong Research Grant Council (CityU 21213626), Guangdong Provincial Natural Science Foundation (Grant No. 2026A1515011937), City University of Hong Kong (Grant No. 9610639).

Data availability. The code and evaluation is available at: <https://github.com/zhangzheng0131/MonoClub>.

Declarations.

Ethics approval. Not applicable.

Consent to participate. Not applicable.

Conflict of interest. The authors declare no conflicts of interest.

Acknowledgements. The authors thank the project contributors and annotators for their valuable work.

References

- [1] Kwon, Y., *et al.*: Golf swing analysis using a wearable motion sensor. *IEEE Sensors Journal* **10**(6), 1081–1082 (2010)
- [2] McNally, W., Vats, K., Pinto, T., Dulhanty, C., McPhee, J., Wong, A.: Golfdb: A video database for golf swing sequencing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2553–2562 (2019). <https://doi.org/10.1109/CVPRW.2019.00311>
- [3] Jung, S., Hong, S., Jeong, J., Jeong, S., Choi, J., Kim, H., Lee, W.: Caddieset: A golf swing dataset with human joint features and ball information. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2025). CVSports Workshop
- [4] Liao, C.-C., Hwang, D.-H., Koike, H.: Ai golf: Golf swing analysis tool for self-training. *IEEE Access* **10**, 106286–106295 (2022) <https://doi.org/10.1109/ACCESS.2022.3210261>
- [5] Ju, C.-Y., Kim, J.-H., Lee, D.-H.: Golfmate: Enhanced golf swing analysis tool through pose refinement network and explainable golf swing embedding for self-training. *Applied Sciences* **13**(20), 11227 (2023) <https://doi.org/10.3390/app132011227>
- [6] Jung, C., Yang, F., Blanchard, N., Wong, H.: Caddie: Compact adaptive detection-driven inference for real-time golf club pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 9978–9987 (2026)
- [7] Dong, L., Yang, Y., Wu, H., Wang, W., Hou, Y., Zhong, Z., Sun, X.: Racketvision: A multiple racket sports benchmark for unified ball and racket analysis. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, pp. 3632–3640 (2026). <https://doi.org/10.1609/aaai.v40i5.37362>
- [8] Urtasun, R., Fleet, D.J., Fua, P.: Monocular 3-d tracking of the golf swing. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 932–938 (2005). <https://doi.org/10.1109/CVPR.2005.229>
- [9] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on*

- Pattern Analysis and Machine Intelligence **43**(1), 172–186 (2021) <https://doi.org/10.1109/TPAMI.2019.2929257>
- [10] Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019). <https://doi.org/10.1109/CVPR.2019.00584>
 - [11] Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., Grundmann, M.: BlazePose: On-Device Real-Time Body Pose Tracking. arXiv:2006.10204 (2020). <https://doi.org/10.48550/arXiv.2006.10204>
 - [12] Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems, vol. 35, pp. 38571–38584 (2022)
 - [13] Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. In: International Journal of Computer Vision, vol. 61, pp. 55–79 (2005)
 - [14] Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
 - [15] Carreira, J., *et al.*: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
 - [16] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (2016)
 - [17] Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision (2018)
 - [18] Lugaresi, C., *et al.*: Mediapipe: A framework for building perception pipelines. In: Workshop on Computer Vision for AR/VR (2019)
 - [19] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28
 - [20] Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P.W., Bauer, U., Menze, B.H.: cldice: A novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16560–16569 (2021). <https://doi.org/10.1109/CVPR46437.2021.01629>
 - [21] Hough, P.V.C.: Method and means for recognizing complex patterns. U.S. Patent

3,069,654 (1962)

- [22] Gioi, R.G., Jakubowicz, J., Morel, J.-M., Randall, G.: Lsd: A line segment detector. *Image Processing On Line* **2**, 35–55 (2012). <https://doi.org/10.5201/ipol.2012.gjmr-lsd>
- [23] Pautrat, R., Barath, D., Larsson, V., Oswald, M.R., Pollefeys, M.: Deeplsd: Line segment detection and refinement with deep image gradients. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17327–17336 (2023). <https://doi.org/10.1109/CVPR52729.2023.01662>
- [24] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems*, vol. 28 (2015)
- [25] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988 (2017). <https://doi.org/10.1109/ICCV.2017.324>
- [26] Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO. GitHub repository (2023). <https://github.com/ultralytics/ultralytics>
- [27] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125 (2017). <https://doi.org/10.1109/CVPR.2017.106>
- [28] Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627–9636 (2019). <https://doi.org/10.1109/ICCV.2019.00972>
- [29] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European Conference on Computer Vision*, pp. 801–818 (2018). https://doi.org/10.1007/978-3-030-01234-2_49
- [30] Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7753–7762 (2019). <https://doi.org/10.1109/CVPR.2019.00794>
- [31] Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131 (2018). <https://doi.org/10.1109/CVPR.2018.00744>

- [32] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics* **34**(6), 248–24816 (2015) <https://doi.org/10.1145/2816795.2818013>
- [33] Cuturi, M., Blondel, M.: Soft-dtw: A differentiable loss function for time-series. In: *Proceedings of the 34th International Conference on Machine Learning*, pp. 894–903 (2017)
- [34] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: *Proceedings of the European Conference on Computer Vision*, pp. 1–21 (2022). https://doi.org/10.1007/978-3-031-20047-2_1
- [35] Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: *Advances in Neural Information Processing Systems*, vol. 30 (2017)
- [36] Liu, P., Wang, J.-H.: Monotrack: Shuttle trajectory reconstruction from monocular badminton video. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022)
- [37] Shen, L., *et al.*: 3d reconstruction of ball trajectory from a single camera in the ball game. In: *International Symposium on Computer Science in Sport* (2015)
- [38] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017)
- [39] Kay, W., *et al.*: The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017)
- [40] Chao, V., *et al.*: 3d ball trajectory reconstruction of a ballistic shot from a monocular basketball video. In: *IEEE IECON* (2023)
- [41] Elfving, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks* **107**, 3–11 (2018)
- [42] Bellman, R.: *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA (1957)