Machine Vision and Applications (2026) 37:137
https://doi.org/10.1007/s00138-026-01895-w

RESEARCH

Broadcast2pitch++: depth-aware game state reconstruction from unconstrained soccer videos

Yewon Hwang[1] · Yin May Oo[1,2] · Muhammad Amrulloh Robbani[1,2] · Vanyi Chao[1,2] · Ankhzaya Jamsrandorj[1] · Hoang Quoc Nguyen[1] · Kyung-Ryoul Mun[1] · Jinwook Kim[1]

Received: 22 May 2026 / Revised: 22 May 2026 / Accepted: 2 August 2026

Abstract

Game State Reconstruction (GSR) aims to recover the spatial positions and semantic identities of all players from broadcast soccer videos, requiring robust localization, tracking, and identity association under unconstrained camera motion and severe visual ambiguity. In this work, we present Broadcast2Pitch++, an extended framework for robust game-state reconstruction from broadcast soccer footage. We introduce two major extensions to the original framework for enhanced temporal identity consistency. First, we propose a depth-aware tracking formulation that incorporates monocular depth cues into the DeepEIoU association process. Second, we propose an improved tracklet refinement strategy that combines appearance similarity, motion consistency, and vision-language-derived identity attributes within a unified soft merge formulation. Comprehensive experiments on the SoccerNet-Tracking and SoccerNet-GSR benchmarks demonstrate that the proposed extensions consistently improve temporal association consistency in challenging broadcast soccer scenarios.

Keywords Game state reconstruction · Multi-object tracking · Sports field registration · VLM

1 Introduction

Game State Reconstruction (GSR) has recently emerged as an important problem in sports analytics, aiming to recover a structured and semantically rich representation of a match from broadcast video. In contrast to conventional multi-object tracking, which primarily focuses on estimating object trajectories, GSR additionally requires identifying players, recovering team affiliations and jersey numbers, and localizing all entities within a pitch coordinate system. Such structured game understanding enables a broad range of downstream applications, including tactical analysis, automated event understanding, performance evaluation, and enhanced fan interaction.

Among team sports, soccer presents a particularly challenging environment for GSR due to rapid player motion, severe and frequent occlusions, strong appearance similarity between same-team players, and unconstrained broadcast camera motion. Recent progress in large-scale public datasets and benchmark initiatives [1–4] has accelerated research toward scalable and automated game understanding systems for broadcast sports analytics. Despite recent advances, reliable player association remains one of the major bottlenecks in practical GSR systems. State-of-the-art

Yin May Oo and Muhammad Amrulloh Robbani contributed equally to this work.

✉ Jinwook Kim
jinwook.kim21@gmail.com

Yewon Hwang
ywhwang@kist.re.kr

Yin May Oo
yinmay@kist.re.kr

Muhammad Amrulloh Robbani
amrullohrobbani@kist.re.kr

Vanyi Chao
vanyi@kist.re.kr

Ankhzaya Jamsrandorj
ankhzaya@kist.re.kr

Hoang Quoc Nguyen
nqhoang@ust.ac.kr

Kyung-Ryoul Mun
krmoon02@kist.re.kr

[1] Intelligence and Interaction Research Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea

[2] AI-Robotics, KIST School, University of Science and Technology, Seoul 02792, Republic of Korea

Published online: 20 August 2026

tracking-by-detection approaches such as DeepEIoU [5] combine motion and appearance cues for data association, yet these cues become highly ambiguous in soccer scenarios where multiple players from the same team exhibit nearly identical visual appearances and frequently overlap in the image plane. Under these conditions, ReID embeddings become less discriminative, while IoU-based motion association alone is often insufficient to prevent identity switches. In addition, tracklet fragmentation caused by occlusion and camera cuts remains a persistent challenge for long-term identity consistency.

In this work, we address these limitations through two major extensions to our previous Broadcast2Pitch framework [6]. First, we introduce a *depth-aware* tracking formulation that integrates monocular depth cues into the DeepEIoU association pipeline. Specifically, we estimate dense inverse-depth maps using Depth-Anything-V2 [7]. The resulting depth cue provides complementary geometric information for distinguishing players that are spatially overlapping and visually similar, thereby improving association robustness in challenging broadcast scenes. Second, we propose an improved tracklet refinement framework that extends our previous ID-Aware Tracklet Refinement (IDATR) strategy [6]. While the original IDATR formulation used vision-language-predicted identity attributes as hard constraints during tracklet merging, the proposed method introduces a soft cost-based formulation that jointly combines appearance similarity, spatial continuity, and graded identity disagreement within a unified merge cost. This formulation improves robustness against noisy attribute predictions caused by occlusion, motion blur, and distant player observations, while preserving the benefits of semantic identity reasoning.

Comprehensive experiments on the SoccerNet-Tracking and SoccerNet-GSR benchmarks demonstrate that the proposed extensions consistently improve tracking association performance over the original Broadcast2Pitch framework and existing baselines.

2 Related work

2.1 Game state reconstruction

Game State Reconstruction (GSR) from broadcast soccer videos aims to estimate the positions and identities of players in a top-view pitch coordinate system. Somers et al. [8] introduced the first end-to-end minimap reconstruction pipeline under the SoccerNet benchmark [2–4, 9–11], combining detection [12–16], tracking, re-identification [17–21], and sports field registration to provide structured game state outputs. Building upon this, Golovkin et al. [22] proposed improvements in camera calibration, spatial alignment, and identity continuity, achieving state-of-the-art performance on the GSR task.

2.2 Multi-object tracking (MOT)

Multi-Object Tracking (MOT) methods generally follow one of two paradigms: tracking-by-detection or end-to-end tracking. Tracking-by-detection decouples tracking into sequential detection and frame-to-frame data association phases. Classical approaches such as DeepSORT [23] combine Kalman-filter-based motion prediction [24] with appearance embeddings for data association, while more recent methods [5, 25–27] improve association robustness through confidence-aware matching strategies and enhanced IoU costs. While highly flexible and modular, these methods are inherently bounded by the performance of the detection model. In contrast, end-to-end tracking methods jointly optimize detection and association within a unified framework, typically using transformer architectures derived from DETR [15]. MOTR [28] introduces an autoregressive query propagation mechanism to update track embeddings continuously over time. To stabilize training and improve initialization, MOTRv2 [29] integrates a YOLOX [14] detection backbone, while MeMOTR [30] embeds memory banks to model long-term temporal dependencies. Although end-to-end models offer elegant, post-processing-free formulations, they generally demand substantially more training data and computational resources than tracking-by-detection approaches.

2.3 Sports field registration

Sports field registration is essential for projecting player positions onto the pitch under challenging broadcast views. TVCalib [31] performs 3D calibration using line and circle reprojection error without keypoints. NBJW [32] directly estimates 3D camera parameters from hierarchical keypoints using DLT and RANSAC in a minimalist geometric setup without refinement. Oo et al. [33] also use DLT and RANSAC for initial homography estimation, but refine it via a refinement network. BroadTrack [34] integrates optical flow with detailed camera models for smooth tracking, while Magera et al. [35] resolve under-constrained views via conic fitting on circular markings. In contrast to these approaches, our work introduces a unified, optimization-based homography framework that robustly integrates dense geometric cues from lines, circles, and keypoints within a single objective, enabling accurate frame-wise estimation without reliance on temporal priors or complex 3D camera models.

2.4 Athlete identification

Consistent identity tracking remains a core challenge in GSR. Traditional appearance-based ReID methods [18, 19] often struggle under occlusions and uniform similarity. Jersey number recognition offers a more robust alternative. Early work by Chan et al. [36] combined CNN-LSTM architectures, while Balaji et al. [37] selected keyframes with clear jersey visibility. Recent approaches employ transformers [38, 39] or vision-language models [17] under weak supervision. PRTreID [20] learns foreground embeddings from body parts and jointly predicts player role and team affiliation through multi-task learning. Other recent approaches [8, 22] use separate models per identity attribute, such as jersey number, role, and team. In contrast, our approach is jointly predict all three identity cues using a single vision-language model. This unified design simplifies the pipeline, removes the need for task-specific modules, and enables more coherent identity reasoning via shared visual-language representations in GSR task.

3 Methodology

Our proposed framework (Fig. 1), Broadcast2Pitch++, reconstructs structured game states from broadcast soccer videos using a modular pipeline. It consists of five key components: (1) player detection and tracking, (2) sports field registration via keypoint and line detection and homography estimation, (3) athlete identification, (4) identity-aware tracklet refinement using a split-and-merge strategy, and (5) post-processing. We adopt YOLOX [14] for detection, DeepEIoU [5] for tracking, which utilizes OSNet [18] ReID features with expansion IoU for improved association. LLaMA−3.2-Vision [40] predicts jersey number, color, and role from player crops. Each module in the pipeline is designed to address specific challenges in broadcast settings, including frequent occlusions, dynamic camera motion, and appearance ambiguity due to similar uniforms.

3.1 Sports field registration

The goal of GSR is to localize athlete positions on a top-view pitch. To achieve this, we estimate a mapping between image-space coordinates and their corresponding pitch coordinates through a sports field registration process. Each athlete's image-space position is approximated by the bottom center of their detected bounding box, which typically corresponds to their foot position on the ground plane. This mapping is modeled by a homography matrix $\mathrm{H} \in \mathbb{R}^{3\times3}$, which transforms image-space coordinates $\mathrm{p}_i \in \mathbb{R}^2$ into pitch coordinates $\mathrm{P}_i \in \mathbb{R}^2$:

$$\mathrm{P}_i = \pi(\mathrm{H} \cdot \tilde{\mathrm{p}}_i), \tag{1}$$

where $\tilde{\mathrm{p}}_i = [x_i, y_i, 1]^\top$ denotes the homogeneous image point, and $\pi(\cdot)$ denotes the projection from homogeneous coordinates to 2D: $\pi([x, y, z]^\top) = [x/z, y/z]^\top$.

We estimate H using a unified optimization that minimizes the residuals between a canonical pitch template of standardized dimensions (eg, 105 m × 68 m) and dense geometric cues–keypoints, lines, and the center circle–detected in the image. In the following sections, we describe the keypoint and line detection network that provides dense geometric cues (Sec. 3.1.1), and then we present the optimization-based homography estimation procedure that integrates line, circle, and fallback keypoint constraints (Sec. 3.1.2).

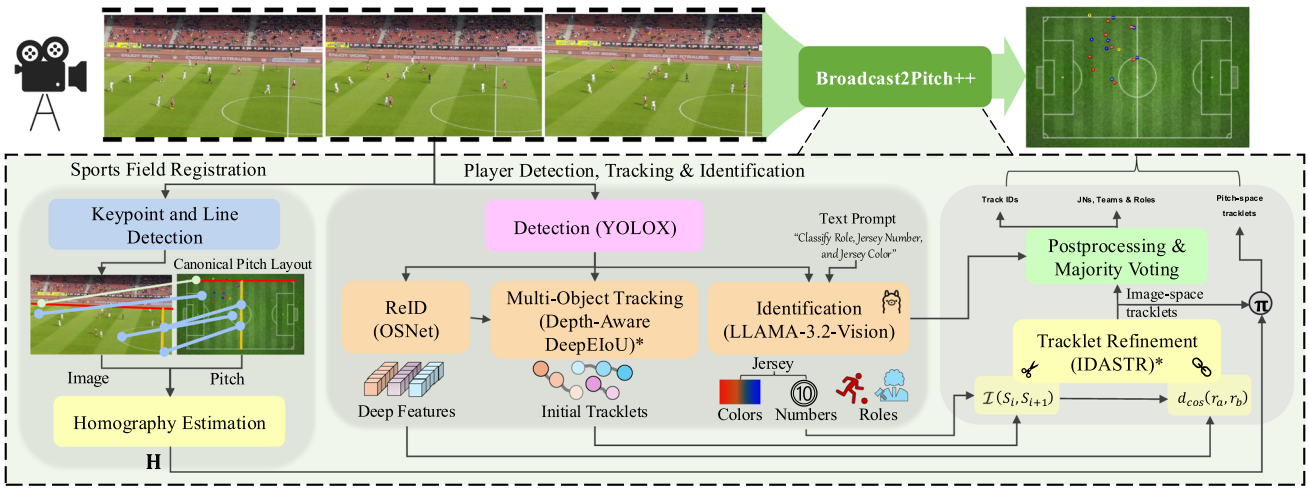


Fig. 1 Overview of our Broadcast2Pitch++ framework. An asterisk (*) indicates components newly modified compared to the original Broadcast-2Pitch framework

3.1.1 Keypoint and line detection model

To predict pitch geometry under broadcast conditions, we adopt a multi-task encoder–decoder built on EfficientNetV2-S with an attention-gated U-Net decoder (Residual EfficientNet–Attention U-Net [33]). The skip connections are modulated by channel–spatial attention gates that suppress irrelevant encoder activations and emphasize strong geometric responses, thereby enhancing spatial sensitivity and reducing redundancy. The network jointly predicts heatmaps for 97 pitch keypoints (Fig. 2a) and 18 pitch lines (Fig. 2b) in [41], and is trained with a heatmap regression approach.

Given an input broadcast frame $I \in \mathbb{R}^{3\times H\times W}$, the network outputs keypoint heatmaps $\mathrm{M}^{\mathrm{kpt}} \in [0,1]^{97\times H\times W}$ and line heatmaps $\mathrm{M}^{\mathrm{line}} \in [0,1]^{18\times H\times W}$. The keypoint maps localize discrete anchors on the field, while the line maps capture extended geometric structure (e.g., sidelines, penalty areas, and the center circle). Used jointly, these two representations provide dense constraints for the homography estimation stage.

3.1.2 Homography estimation

In the homography estimation step, we solve for the inverse transformation $\mathrm{G} = \mathrm{H}^{-1}$, which maps points from the pitch template to the image space, allowing us to compute residuals directly where the keypoints and lines are detected.

Line constraints

For each pitch line excluding center circle, L (Fig. 2b), we extract high-confidence samples $\{(x_{L,j}, y_{L,j}, w_{L,j})\}_{j=1}^{N}$ from its predicted heatmap, $\mathrm{M}^{\mathrm{line}}$ in Sec. 3.1.1, by thresholding ($\tau > 0.8$) and randomly sampling up to N points from the remaining high-intensity pixels; each sample retains its heatmap value as a confidence weight $w_{L,j}$.

Let ℓ_L be the line L expressed in homogeneous coordinates in the canonical pitch template. This line is projected to the image space as $\ell'_L \sim \mathrm{G}^{\top}\ell_L$. For each image-space sample $\tilde{\mathrm{p}}_{L,j} = [x_{L,j}, y_{L,j}, 1]^{\top}$, we compute a confidence-weighted algebraic line residual:

$$r^{\mathrm{line}}_{L,j}(\mathrm{G}) = w_{L,j} \cdot {\ell'_L}^{\top}\tilde{\mathrm{p}}_{L,j}, \tag{2}$$

which measures how well the sampled point aligns with the projected line.

Circle (conic) constraint

Let C denote the center circle on the pitch, centered at $\mathrm{o} = (x_0, y_0, 1)^{\top}$ with radius r (e.g., 9.15m). In the pitch template space, this circle is represented as a conic, defined by the matrix:

$$\mathrm{C} = \begin{bmatrix} 1 & 0 & -x_0 \\ 0 & 1 & -y_0 \\ -x_0 & -y_0 & x_0^2 + y_0^2 - r^2 \end{bmatrix}, \tag{3}$$

which corresponds to the homogeneous form of the 2D circle equation $(x - x_0)^2 + (y - y_0)^2 = r^2$, expressed as $\mathrm{x}^{\top}\mathrm{C}\mathrm{x} = 0$ for $\mathrm{x} = [x, y, 1]^{\top}$.

Under homography, the conic transforms as $\mathrm{C}' = \mathrm{H}^{-\top}\mathrm{C}\mathrm{H}^{-1} = \mathrm{G}^{\top}\mathrm{C}\mathrm{G}$. Similar to *line constraints*, we extract high-confidence samples $\tilde{\mathrm{p}}_k = [x_k, y_k, 1]^{\top}$ from the detected center circle heatmap, $\mathrm{M}^{\mathrm{line}}$, each with confidence weight w_k. We compute the confidence-weighted algebraic conic residual:

$$r^{\mathrm{circ}}_k(\mathrm{G}) = w_k \cdot \tilde{\mathrm{p}}_k^{\top}\mathrm{C}'\tilde{\mathrm{p}}_k. \tag{4}$$

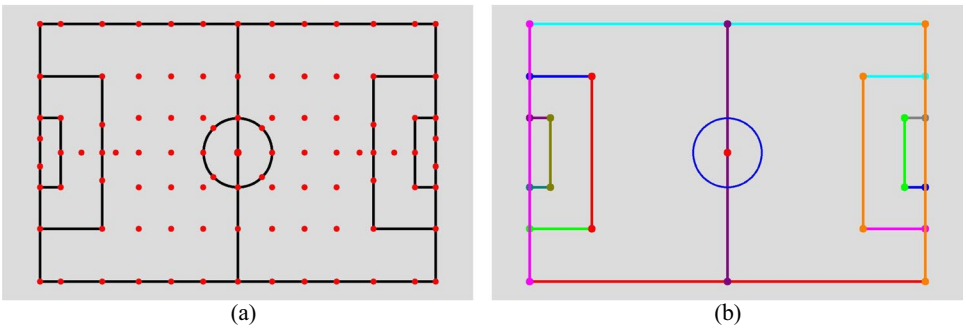
Keypoint fallback (when lines are sparse)

When the number of detected lines is insufficient, we incorporate confidence-weighted point correspondences from predicted keypoint heatmaps, $\mathrm{M}^{\mathrm{kpt}}$. Let $(\mathrm{p}_i, \mathrm{P}_i, c_i)$ denote an image–template keypoint correspondence with confidence score c_i. The confidence-weighted reprojection residual is:

$$r^{\mathrm{kpt}}_i(\mathrm{G}) = c_i \cdot \left\| \pi(\mathrm{G}^{-1}\mathrm{p}_i) - \mathrm{P}_i \right\|_2, \tag{5}$$

Homography objective and solver

We estimate G by minimizing the least-squares objective of lines, circle, and keypoints:



(a) (b)

Fig. 2 Canonical pitch layout with keypoint and line representations.**a** The 97 pitch keypoints (red dots) serve as discrete anchors for localization. **b** The 18 pitch lines (colored polylines) capture extended geometric structures, including sidelines, penalty areas, and the center circle. Together, these representations provide dense geometric cues for robust homography estimation

$$\min_{\mathrm{G}} \underbrace{\sum_{L}\sum_{j}\left(r_{L,j}^{\mathrm{line}}(\mathrm{G})\right)^2}_{\text{Lines}} + \underbrace{\sum_{k}\left(r_{k}^{\mathrm{circ}}(\mathrm{G})\right)^2}_{\text{Circle}} + \underbrace{\sum_{i}\left(r_{i}^{\mathrm{kpt}}(\mathrm{G})\right)^2}_{\text{Keypoints}} \tag{6}$$

We initialize G using a DLT [42] estimate from at least four keypoint correspondences (when available), and otherwise default to the identity matrix; then refine it using Levenberg–Marquardt algorithm. The hybrid line–conic terms anchor the global pitch geometry, while the keypoint fallback improves robustness in cases with sparse line visibility.

3.2 Depth-aware tracking

We utilize DeepEIoU [5] as the baseline tracker, which combines an Expanded-IoU motion cue with an OSNet-based ReID appearance cue. Although the two cues are largely complementary, they share a common failure mode in broadcast soccer footage: when two players from the same team move close to each other in the image plane, multiple candidate associations may exhibit similarly high IoU scores, while their ReID embeddings become nearly indistinguishable due to identical jerseys.

To address this issue, we introduce a third cue based on monocular depth. Even when two players overlap in image space and exhibit highly similar visual appearances, they are unlikely to occupy the same relative depth with respect to the camera. Depth therefore serves as a strong disambiguation cue precisely in situations where IoU and ReID become unreliable.

In this work, we employ Depth-Anything-V2-Small [7] as the monocular depth estimator, which produces a dense inverse-depth map for each frame.

3.2.1 Depth feature for detections

Let $\mathcal{D}_t : \mathbb{R}^2 \to \mathbb{R}_+$ denote the dense inverse-depth map at frame t. For a detection j with bounding box $\mathrm{b}_j = (x_1, y_1, x_2, y_2)$, we define the foot sub-region as the lower half of the bounding box,

$$\mathcal{B}_j^{\mathrm{foot}} = \left\{(x,y) \mid x_1 \le x \le x_2,\ \tfrac{y_1+y_2}{2} \le y \le y_2\right\}, \tag{7}$$

which approximately corresponds to the player's lower-body and ground-contact area. This region is the more stable region for player depth estimation under partial upper-body occlusion.

To obtain a robust per-detection depth estimate, we compute the spatial median of the inverse-depth values within $\mathcal{B}_j^{\mathrm{foot}}$,

$$\tilde{d}_j = \underset{(x,y)\in\mathcal{B}_j^{\mathrm{foot}}}{\mathrm{median}} \mathcal{D}_t(x,y). \tag{8}$$

which reduces sensitivity to local texture noise and outlier depth predictions.

Because Depth-Anything-V2 predicts relative rather than metric depth, the absolute scale of $\tilde{d}_j$ may vary across frames. We therefore normalize the raw inverse-depth estimates within each frame using the global minimum and maximum inverse-depth values of that frame, yielding a normalized depth value $d_j \in [0,1]$:

$$d_j = \frac{\tilde{d}_j - \min_{(x,y)} \mathcal{D}_t(x,y)}{\max_{(x,y)} \mathcal{D}_t(x,y) - \min_{(x,y)} \mathcal{D}_t(x,y) + \varepsilon}. \tag{9}$$

The resulting scalar encodes the relative camera-depth ordering of the detection within the current frame.

3.2.2 Depth state for tracklets

Each active tracklet k maintains a temporally smoothed depth estimate d_k, which is updated after every successful association using an exponential moving average (EMA),

$$d_k \leftarrow \alpha_d\, d_j + (1-\alpha_d)\, d_k, \qquad \alpha_d = 0.5, \tag{10}$$

where d_j denotes the normalized depth value of the newly associated detection. Newly activated tracks are initialized with $d_k = d_j$.

3.2.3 Pairwise depth cost

For the set of N active tracklets and M detections in the current frame, we construct a depth-based association cost matrix $\mathrm{C}^{\mathrm{depth}} \in [0,1]^{N\times M}$, defined as

$$C_{kj}^{\mathrm{depth}} = \min\left(1,\ \frac{|d_k - d_j|}{\sigma_d}\right), \qquad \sigma_d = 0.25, \tag{11}$$

where d_k and d_j denote the normalized depth values of tracklet k and detection j, respectively. Consequently, $C_{kj}^{\mathrm{depth}} = 0$ when the two depth estimates coincide, and saturates to 1 once their difference exceeds σ_d.

3.2.4 Baseline DeepEIoU cost

For completeness, we briefly summarize the original DeepEIoU [5] association cost formulation before introducing the proposed depth-aware extension. Let $\mathrm{expand}(\mathrm{b}, e)$ denote the bounding box b expanded by ratio e along each

side. The expanded-IoU cost between tracklet k and detection j is defined as

$$C_{kj}^{\mathrm{eIoU}} = 1 - \mathrm{IoU}(\mathrm{expand}(\mathrm{b}_k, e),\ \mathrm{expand}(\mathrm{b}_j, e)), \tag{12}$$

where lower values indicate stronger spatial agreement.

Appearance similarity is measured using the cosine distance between L2-normalized OSNet embeddings,

$$C_{kj}^{\mathrm{ReID}} = \frac{1}{2}\left(1 - \cos(\mathrm{f}_k, \mathrm{f}_j)\right). \tag{13}$$

The appearance cost is further gated using spatial and appearance thresholds τ_{prox} and τ_{app},

$$\tilde{C}_{kj}^{\mathrm{ReID}} = \begin{cases} 1, & C_{kj}^{\mathrm{eIoU}} > \tau_{\mathrm{prox}}, \\ 1, & C_{kj}^{\mathrm{ReID}} > \tau_{\mathrm{app}}, \\ C_{kj}^{\mathrm{ReID}}, & \text{otherwise}, \end{cases} \tag{14}$$

thereby suppressing candidate pairs that are either spatially distant or appearance-inconsistent.

The baseline DeepEIoU association cost is then computed as

$$C_{kj}^{\mathrm{base}} = \min\left(C_{kj}^{\mathrm{eIoU}},\ \tilde{C}_{kj}^{\mathrm{ReID}}\right), \tag{15}$$

which retains the more permissive of the motion and appearance cues for each candidate pair.

3.2.5 Depth-fused matching cost

To incorporate monocular depth information into the association process, we combine the baseline cost C^{base} with the proposed depth cost C^{depth} through a weighted average,

$$C_{kj} = \frac{w_b\, C_{kj}^{\mathrm{base}} + w_d\, C_{kj}^{\mathrm{depth}}}{w_b + w_d}, \qquad w_b = 1.0,\ w_d = 0.3. \tag{16}$$

A weighted-average formulation is adopted instead of a hard minimum-based fusion for two reasons. First, the fused cost remains bounded within the interval [0, 1], thereby preserving compatibility with the original DeepEIoU matching threshold τ_{match}. Second, the formulation enables depth to operate as a soft regularization cue rather than a hard veto. Specifically, even a strong depth disagreement ($C_{kj}^{\mathrm{depth}} \approx 1$) contributes at most $w_d/(w_b + w_d) \approx 0.23$ to the final cost of an otherwise highly confident match. As a result, the depth term is insufficient to independently reject a valid association, yet remains capable of disambiguating borderline cases in which motion and appearance cues alone are ambiguous.

Final assignment is performed using the Hungarian algorithm on the fused association matrix C,

$$\hat{\pi} = \arg\min_{\pi} \sum_{(k,j)\in\pi} C_{kj} \quad \text{s.t.} \quad C_{kj} \le \tau_{\mathrm{match}}, \tag{17}$$

where $\tau_{\mathrm{match}} = 0.8$ is identical to the threshold used in the original DeepEIoU implementation.

3.2.6 Application to the multi-step association cascade

DeepEIoU [5] performs data association using a three-stage cascade strategy at each frame: (i) association with high-confidence detections using larger expansion ratios to improve tolerance to motion and localization uncertainty, (ii) association with low-confidence detections using a smaller expansion ratio and without the ReID term (i.e., $C_{kj}^{\mathrm{base}} = C_{kj}^{\mathrm{eIoU}}$) to suppress false associations caused by unreliable detections and noisy appearance features, and (iii) association of unconfirmed tracklets using smaller expansion ratios together with a relaxed matching threshold to facilitate reliable early track formation while reducing erroneous associations.

The proposed fused cost in Eq. (16) is applied consistently across all stages of the cascade using the same weighting parameters (w_b, w_d), allowing the depth cue to contribute uniformly throughout the entire association process.

3.3 Athlete identification

Athlete identification is a core component of GSR, involving the recognition of roles, jersey numbers, and team affiliations. These identity cues are critical in sports scenarios where players often appear visually similar, making traditional appearance-based tracking ambiguous [5, 23]. We formulate identity recognition (jersey number, color, and role) as a multimodal autoregressive generation task. Specifically, we leverage LLaMA−3.2-Vision [40], an instruction-tuned vision-language model, to estimate the conditional likelihood of a sequence of identification tokens given both a natural language prompt and visual embeddings extracted from cropped athlete bounding boxes:

$$P(x \mid q, v) = \prod_{t=1}^{T} P(x_t \mid x_{<t}, q, v) \tag{18}$$

where $x = (x_1, x_2, \ldots, x_T)$ denotes the sequence of identity tokens, q is the instruction prompt ("Classify role, jersey number, and jersey color"), and v represents the visual embedding produced by the model's visual encoder.

Instruction-tuned VLMs offer strong visual grounding and language understanding, offering several key advantages for athlete identification: 1) improved generalization to unseen identity attributes (e.g., novel jersey colors), which is essential for open-set recognition, 2) reduced annotation and engineering overhead, as prompt tuning enables flexible multi-task inference without requiring architectural modifications, and 3) scalability and adaptability across a wide range of sports environments and game formats.

3.4 ID-aware soft tracklet refinement (IDASTR)

Following the initial track generation stage using Deep-EIoU [5], we further refine the resulting trajectories through a split-and-merge strategy derived from our previously proposed ID-Aware Tracklet Refinement (IDATR) framework [6], which we term *ID-Aware Soft Tracklet Refinement (IDASTR)*. The original IDATR leverages frame-wise identity attributes predicted by a vision-language model (VLM), including jersey number, jersey color, and player role, to improve the discrimination of visually similar players. Although these identity cues are highly informative in sports tracking scenarios, per-detection VLM predictions are often unreliable under severe occlusion, motion blur, or distant broadcast viewpoints, leading to sporadic attribute inconsistencies within the same trajectory.

In the original IDATR formulation, identity disagreement was treated as a hard constraint during the merge stage, such that a mismatch in any majority-voted attribute immediately vetoed a candidate merge. While effective under reliable attribute predictions, this formulation is sensitive to local prediction noise and may incorrectly reject otherwise valid associations. To address this limitation, we introduce a soft merge formulation in which appearance similarity, spatial continuity, and identity disagreement are integrated into a single fused association cost.

Split The split stage remains identical to the original IDATR formulation. Identity switches typically arise when a tracklet contains a large temporal gap, reflecting a player's exit and re-entry. To address this, we split tracklets where the frame gap exceeds a threshold τ_{gap}. Let $F = [f_1, f_2, ..., f_N]$ denote frame indices where the tracklet appears, and let $\Delta_f = f_{i+1} - f_i$ denote the temporal gap between consecutive frames. We define the candidate split points as $G = \{g \mid \Delta_f > \tau_{gap}\}$, and segment the tracklet into $[S_1, S_2, ..., S_{|G|+1}]$. Adjacent segments S_i and S_{i+1} are then evaluated for identity inconsistency using the binary indicator

$$\mathcal{I}(X,Y) = \begin{cases} 1 & \begin{array}{l}\text{if } \mathrm{Maj}(R_X) \neq \mathrm{Maj}(R_Y) \\ \vee\ \mathrm{Maj}(J_X) \neq \mathrm{Maj}(J_Y) \\ \vee\ \mathrm{Maj}(C_X) \neq \mathrm{Maj}(C_Y)\end{array} \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

where $\mathrm{Maj}(R_X)$, $\mathrm{Maj}(J_X)$, and $\mathrm{Maj}(C_X)$ denote the majority-voted role, jersey number, and color for X. A split is enforced between S_i and S_{i+1} whenever $\mathcal{I}(S_i, S_{i+1}) = 1$.

Merge After tracklet splitting, candidate tracklet pairs are merged based on a fused cost that combines appearance similarity, spatial continuity, and identity consistency.

Appearance cost Appearance similarity is measured using the cosine distance between the mean ReID embeddings of two tracklets,

$$d_{cos}(r_a, r_b) = 1 - \frac{r_a \cdot r_b}{\|r_a\|\|r_b\|}. \tag{20}$$

Spatial-continuity residual Let p_a^{end} and p_b^{start} denote the bottom-center image coordinates at the end of tracklet T_a and the beginning of T_b, respectively, and let $\Delta f_{a,b}$ represent the temporal gap between them. Assuming a maximum plausible image-plane displacement $v_{\max}$, the maximum feasible displacement over the temporal gap is $v_{\max}\Delta f_{a,b}$. We therefore define the normalized traversal cost as

$$c_{trav}(T_a, T_b) = \min\left(1,\ \frac{\|p_a^{end} - p_b^{start}\|_2}{v_{\max}\Delta f_{a,b}}\right). \tag{21}$$

To avoid penalizing physically plausible transitions, only the excess traversal cost beyond the threshold τ_{trav} contributes to the final merge cost,

$$r_{trav}(T_a, T_b) = \max\big(0,\ c_{trav}(T_a, T_b) - \tau_{trav}\big). \tag{22}$$

Identity-disagreement residual. Unlike the original IDATR [6], which used a binary identity constraint, we employ a soft identity disagreement term defined as

$$\overline{\mathcal{I}}(T_a, T_b) = \frac{1}{3}\begin{pmatrix} 1[Maj(R_a) \neq Maj(R_b)] \\ +1[Maj(J_a) \neq Maj(J_b)] \\ +1[Maj(C_a) \neq Maj(C_b)] \end{pmatrix}, \tag{23}$$

which measures the fraction of disagreeing identity attributes between two tracklets. The residual equals 0 when all attributes agree and 1 when all attributes disagree.

Fused merge cost The three component terms are combined into a unified merge cost,

$$\begin{aligned} &\mathcal{C}(T_a, T_b) \\ &= d_{cos}(r_a, r_b) + w_{trav} r_{trav}(T_a, T_b) + w_{attr}\overline{\mathcal{I}}(T_a, T_b), \end{aligned} \tag{24}$$

where w_{trav} and w_{attr} control the relative contribution of the spatial and identity terms, respectively.

Two tracklets are merged when: (i) their fused cost falls below a threshold τ_{merge}, (ii) their temporal intervals do not overlap, and (iii) the required spatial traversal remains physically feasible,

$$\mathcal{M}(T_a, T_b) = \begin{cases} 1 & \text{if } \mathcal{C}(T_a, T_b) < \tau_{merge} \\ & \wedge\ F_a \cap F_b = \emptyset \\ & \wedge\ c_{trav}(T_a, T_b) < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

Candidate pairs are iteratively merged in ascending order of $\mathcal{C}$ until no valid merge candidates remain. We use $\tau_{gap} = 25$, $\tau_{merge} = 0.45$, $\tau_{trav} = 0.5$, $v_{\max} = 80$ px/frame, $w_{trav} = 0.5$, and $w_{attr} = 0.2$ throughout all experiments.

Comparison to the Original IDATR [6] **.** The proposed formulation differs from the original IDATR in two key aspects. First, the binary identity indicator $\mathcal{I}(T_a, T_b)$, which previously acted as a hard rejection criterion, is replaced with the soft identity disagreement term $\bar{\mathcal{I}}(T_a, T_b)$ [6]. Consequently, disagreement in a single attribute no longer immediately rejects a merge that is otherwise strongly supported by appearance similarity and motion cues. Second, the original fixed spatial threshold $\tau_{spatial}$ [6] is replaced by the normalized residual r_{trav}, which lies on the same [0, 1] scale as the remaining cost terms. This allows appearance similarity, spatial continuity, and identity consistency to compete within a unified fused cost formulation governed by a single threshold τ_{merge}.

3.5 Post-processing

After obtaining tracklets from IDATR and LLaMA−3.2-Vision [40], team affiliation (left/right) is inferred by clustering jersey colors and comparing the mean spatial *x*-position of each cluster – assuming players tend to stay closer to their own half. While dynamic game scenarios such as high pressing or set pieces can momentarily obscure spatial distinctions between teams, we argue that players generally exhibit a persistent positional bias toward their own defensive half. Consequently, even during dynamic phases of play, the average *x*-positions of players from opposing teams tend to remain meaningfully separated over time, as offensive advances are typically met with coordinated defensive retreats that preserve this spatial distinction. A final post-processing stage improves identity consistency and team-level coherence. This includes majority voting within each tracklet for jersey number, role, and team affiliations, and filtering unreliable tracklets (e.g., tracklets with length less than 20 and detections outside of the soccer field in pitch coordinate) to ensure robust reconstruction.

4 Experiments

4.1 Dataset

We evaluate our GSR framework on the SoccerNet-GSR dataset [8], which contains 200 broadcast video clips, each 30 s long. It extends SoccerNet-Tracking [3] and provides dense geometric annotations over 9.37 million line points for pitch localization and camera calibration, as well as over 2.36 million athlete positions on the pitch, role, team, and jersey number. The dataset is split into 57 training, 58 validation, 49 test, and 36 challenge clips.

4.2 Evaluation metrics

GSR We evaluate our GSR results using GS-HOTA [8], a metric specifically designed to assess GSR by jointly measuring detection accuracy, identity association, and localization accuracy. Formally, GS-HOTA is computed as follows:

$$GS-HOTA(P, G) = LocSim(P, G) \times IdSim(P, G), \tag{26}$$

where

$$LocSim(P, G) = e^{\ln(0.05)\frac{||P-G||_2^2}{r^2}}, \tag{27}$$

and

$$IdSim(P, G) = \begin{cases} 1 & \text{if all ID attributes match,} \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

LocSim(P, G) quantifies the spatial proximity between predicted and ground-truth tracklets in pitch coordinates, while *IdSim*(P, G) penalizes the framework's performance if any of the three identity attributes–role, jersey number, or team–are misclassified. This highlights the importance of not only accurate tracking but also comprehensive identity preservation across all attribute dimensions. GS-HOTA uses a 5-meter spatial tolerance to account for localization errors in pitch space. As the primary evaluation metric for the SoccerNet-GSR benchmark, GS-HOTA provides a holistic view of system performance. In addition to the GS-HOTA, we report its sub-metrics–GS-DetA (detection accuracy), GS-AssA (association accuracy), and IDF1–to quantify our approach's effectiveness in detecting all relevant subjects and maintaining consistent identities across time.

Multi-Object Tracking Following the evaluation protocol of SoccerNet-Tracking [3], we evaluate the impact of the proposed depth-aware association strategy on multi-object tracking performance using four standard MOT metrics: Higher Order Tracking Accuracy (HOTA), Detection

Table 1 Tracking performance on the SoccerNet-Tracking test set. The suffix "-ft" indicates models fine-tuned on the SoccerNet-Tracking training set

Algorithm	HOTA↑	DetA↑	AssA↑	MOTA↑
Method: End-to-End				
MeMOTR-ft [30]	59.40	65.93	53.64	81.10
MOTR-ft [28]	55.06	58.12	52.28	66.4
Method: Tracking-by-Detection				
DeepSORT [23]	36.66	40.02	33.76	33.91
ByteTrack [25]	47.23	44.49	50.26	31.74
FairMOT-ft [43]	57.88	66.57	50.49	83.57
DeepEIoU [5]	85.98	**99.79**	74.07	99.44
DeepEIoU+Depth(Ours)	**86.54**	99.76	**75.07**	**99.71**

The best results across all compared architectures are highlighted in bold

Accuracy (DetA), Association Accuracy (AssA), and Multiple Object Tracking Accuracy (MOTA). HOTA jointly measures detection and association quality, while DetA and AssA separately quantify detection precision and temporal association consistency, respectively. MOTA evaluates overall tracking performance by accounting for false positives, missed detections, and identity switches.

4.3 Implementation details

We employ YOLOX [14], trained on the SoccerNet-Tracking dataset [3], and OSNet [18], trained on the SportsMOT dataset [1], for player detection and re-identification, respectively. Our keypoint and line detection model is trained for 25 epochs with a learning rate of $1e^{-4}$ on the SoccerNet-GSR [8] train set, each resized to 384×384 pixels. The LLaMA−3.2-Vision model is fine-tuned [40] on the SoccerNet jersey dataset [9] for just one epoch with a learning rate of $1e^{-5}$. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24 GB memory using the PyTorch framework.

4.4 Quantitative evaluation

Multi-Object Tracking To isolate the impact of the proposed depth-aware association on multi-object tracking performance, we evaluate the method on the SoccerNet-Tracking test set in image space and compare it against existing tracking algorithms. As shown in Table 1, end-to-end tracking approaches exhibit comparatively weaker performance in the soccer domain, likely due to the complex camera motion, frequent occlusions, and dense player interactions present in broadcast soccer footage. Among the tracking-by-detection methods, DeepEIoU achieves the strongest overall performance, demonstrating the effectiveness of combining motion and appearance cues for sports tracking scenarios. By incorporating the proposed depth-aware association cue into DeepEIoU, tracking performance is further improved across nearly all evaluation metrics. In particular, the proposed method achieves higher HOTA, AssA, and MOTA scores compared with the original DeepEIoU baseline, indicating improved overall tracking quality and temporal association consistency. A marginal decrease is observed only in DetA, suggesting that the depth cue primarily benefits association robustness rather than detection quality itself.

Table 2 Game-state reconstruction tracking performance after projection onto pitch coordinates using homography estimation with identity attributes disabled

Config	GS-HOTA	GS-DetA	GS-AssA	IDF1
DeepEIoU+IDATR (baseline)	79.52	85.97	73.60	83.40
DeepEIoU+Depth+IDATR	79.82	84.57	75.36	83.46
DeepEIoU+IDASTR	80.91	**86.12**	76.05	85.84
DeepEIoU+Depth+IDASTR	**81.18**	84.58	**77.94**	**85.93**

The best results across all compared configurations are highlighted in bold

GSR We evaluate the effectiveness of our Broadcast2Pitch++ framework on SoccerNet-GSR benchmark. Table 2 evaluates the impact of the proposed depth-aware tracking and tracklet refinement strategies on GS-HOTA, GS-DetA, GS-AssA, and IDF1 without player ID attributes (e.g., jersey number, role, team). As shown in the results, introducing the proposed depth-aware association and IDASTR consistently improves game-state reconstruction tracking performance. In particular, the largest gains are observed in GS-AssA and IDF1, indicating improved association consistency and identity preservation after projection onto pitch coordinates. Incorporating monocular depth into the DeepEIoU association process improves GS-AssA from 73.60 to 75.36, demonstrating that depth cues provide complementary geometric information for resolving ambiguities among visually similar and spatially overlapping players. Furthermore, replacing the original hard-constrained IDATR with the proposed soft cost-based refinement strategy (IDASTR) yields additional improvements across all tracking metrics, demonstrating its efficacy. Combining both depth-aware association and IDASTR achieves the best overall performance, improving GS-HOTA from 79.52 to 81.18 and GS-AssA from 73.60 to 77.94. These results demonstrate that the proposed extensions are particularly effective at improving identity consistency in challenging broadcast soccer scenarios.

Table 3 presents the quantitative results on the SoccerNet-GSR test set with player ID attributes, comparing our method against the previous GSR frameworks [6, 8, 22]. Our best-performing configuration (DeepEIoU+IDASTR) achieves new state-of-the-art performance across all evaluation metrics, obtaining a GS-HOTA score of 62.56. In

Table 3 Performance comparison on the SoccerNet-GSR test set. We compare our framework with other baselines with identity attributes enabled

Framework	GS-HOTA ↑	GS-DetA ↑	GS-AssA ↑	IDF1 ↑
GSR Baseline [8]	22.26	10.67	46.46	–
Golovkin et al. [22]	55.82	41.67	74.86	–
Broadcast2Pitch [33]	61.48	48.47	78.00	64.20
DeepEIoU+Depth+IDATR	57.91	42.07	79.72	58.7
DeepEIoU+IDASTR (our best)	**62.56**	**48.85**	80.12	**65.06**
DeepEIoU+Depth+IDASTR	58.91	42.40	**81.87**	59.60

The best results across all compared architectures on the SoccerNet-GSR test set are highlighted in bold

particular, all newly proposed variants consistently improve GS-AssA compared with previous baselines, demonstrating the effectiveness of the proposed depth-aware tracking and soft tracklet refinement strategies for improving temporal association consistency.

The comparatively lower performance of DeepEIoU+Depth+IDATR and DeepEIoU+Depth+IDASTR is primarily attributed to reductions in GS-DetA. This behavior arises from two factors. First, incorporating depth introduces an additional association constraint during tracking, which leads to more conservative matching and consequently increases the number of missed detections and fragmented trajectories. Second, the resulting short fragmented tracklets are more susceptible to noisy majority-vote identity attributes under imperfect VLM predictions, which negatively affects attribute-aware evaluation. As a result, the gains in association quality are partially offset by reductions in detection-related metrics.

Nevertheless, GS-AssA consistently improves under both ID attribute-enabled and -disabled evaluation settings, indicating that the proposed depth-aware association effectively improves temporal identity consistency independently of the evaluation protocol. Specifically, introducing the proposed depth-aware association into DeepEIoU improves GS-AssA by $+1.76$ under attrs-OFF evaluation and by $+1.72$ under attrs-ON evaluation, while decreases GS-DetA by -1.40 and -6.40, respectively. These results suggest that the proposed method primarily benefits the association robustness rather than detection.

5 Conclusion

We present Broadcast2Pitch++, an extended framework for Game State Reconstruction (GSR) from unconstrained broadcast soccer videos. Building upon our previous Broadcast2Pitch framework, the proposed method introduces two major extensions for improving temporal association consistency under challenging broadcast soccer scenarios. First, we incorporate monocular depth cues into the DeepEIoU tracking pipeline through a depth-aware association strategy that leverages dense inverse-depth maps to disambiguate visually similar and spatially overlapping players. Second, we propose ID-Aware Soft Tracklet Refinement (IDASTR), a refined split-and-merge optimization framework that combines appearance similarity, motion consistency, and soft identity disagreement into a unified tracklet association cost. Comprehensive experiments on the SoccerNet-Tracking and SoccerNet-GSR benchmarks demonstrate that the proposed extensions consistently improve tracking association quality. In this regard, Broadcast2Pitch++ provides useful insights into the role of complementary cues such as depth and soft identity-aware association for improving tracking consistency in real-world sports analytics systems.

6 Limitations and future work

Despite the improvements introduced in Broadcast2Pitch++, several limitations remain. First, the vision-language model used for identity attribute prediction, LLaMA−3.2-Vision, is computationally expensive and currently unsuitable for real-time deployment in practical broadcast sports analytics systems. Future work will therefore investigate lighter vision-language architectures that can achieve comparable identity recognition performance while significantly reducing computational overhead and inference latency. Second, while the proposed depth-aware association improves temporal association consistency, it also introduces a tradeoff between association quality and detection-related metrics. In particular, the additional depth constraint may overly suppress borderline associations, leading to increased track fragmentation and reduced detection performance under certain evaluation settings. Future work will focus on more adaptive depth fusion strategies and improved confidence-aware association mechanisms to better balance association robustness and detection accuracy.

Acknowledgements This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2026-25516375, Development of Multimodal Sensory Sharing Technology across Heterogeneous XR Devices).

Author contributions Jinwook Kim: Conceptualization, Funding acquisition, Project administration, Resources, Supervision. Yewon Hwang: Writing original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Yin May Oo: Writing original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Muhammad Amrulloh Robbani: Writing original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Vanyi Chao: Formal analysis, Validation. Ankhzaya Jamsrandorj: Formal analysis, Validation. Hoang Quoc Nguyen: Formal analysis, Valida-

tion. Kyung-Ryoul Mun: Supervision, Review & editing.

Data availibility SoccerNet GSR and SoccerNet Tracking dataset are available at the following URL: https://huggingface.co/datasets/SoccerNet/SN-GSR-2025 and https://github.com/SoccerNet/sn-tracking, respectively.

References

1. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: Sportsmot: a large multi-object tracking dataset in multiple sports scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9921–9931. (2023)
2. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: a scalable dataset for action spotting in soccer videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1711–1721. (2018)
3. Cioppa, A., Giancola, S., Deliege, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., Van Droogenbroeck, M.: Soccernet-tracking: multiple object tracking dataset and benchmark in soccer videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3491–3502. (2022)
4. Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasrollahi, K., Ghanem, B., Moeslund, T.B., Van Droogenbroeck, M.: Soccernet-v2: a dataset and benchmarks for holistic understanding of broadcast soccer videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4508–4519. (2021)
5. Huang, H.-W., Yang, C.-Y., Sun, J., Kim, P.-K., Kim, K.-J., Lee, K., Huang, C.-I., Hwang, J.-N.: Iterative scale-up expansioniou and deep features association for multi-object tracking in sports. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 163–172 (2024)
6. Oo, Y.M., Hwang, Y., Robbani, M.A., Chao, V., Jamsrandorj, A., Nguyen, H.Q., Mun, K.-R., Kim, J.: Broadcast2pitch: game state reconstruction from unconstrained soccer videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4483–4493. (2026)
7. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv:2406.09414 (2024)
8. Somers, V., Joos, V., Cioppa, A., Giancola, S., Ghasemzadeh, S.A., Magera, F., Standaert, B., Mansourian, A.M., Zhou, X., Kasaei, S.: Soccernet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3293–3305. (2024)
9. Cioppa, A., Deliège, A., Giancola, S., Ghanem, B., Droogenbroeck, M.V.: Scaling up soccernet with multi-view spatial localization and re-identification. Sci. Data**9**, (2022)
10. Cioppa, A., Deliege, A., Magera, F., Giancola, S., Barnich, O., Ghanem, B., Van Droogenbroeck, M.: Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4537–4546. (2021)
11. Mkhallati, H., Cioppa, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M.: Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5074–5085. (2023)
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inform. Process. Syst. **28**, (2015)
13. Reis, D., Kupec, J., Hong, J., Daoudi, A.: Real-time flying object detection with yolov8, (2023). arXiv:2305.09972
14. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021, (2021). arXiv:2107.08430
15. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229. Springer (2020)
16. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965–16974. (2024)
17. Habel, K., Deuser, F., Oswald, N.: Clip-reident: Contrastive training for player re-identification. In: Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports, pp. 129–135. (2022)
18. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3702–3712. (2019)
19. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning generalisable omni-scale representations for person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 5056–5069 (2021)
20. Mansourian, A.M., Somers, V., De Vleeschouwer, C., Kasaei, S.: Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking. In: Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports, pp. 103–112. (2023)
21. Somers, V., De Vleeschouwer, C., Alahi, A.: Body part-based representation learning for occluded person re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1613–1623. (2023)
22. Golovkin, V., Nemtsev, N., Shandyba, V., Udin, O., Kasatkin, N., Kononov, P., Afanasiev, A., Ulasen, S., Boiarov, A.: From broadcast to minimap: achieving state-of-the-art soccernet game state reconstruction, (2025). arXiv:2504.06357
23. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649. IEEE (2017)
24. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 3464–3468 (2016). https://doi.org/10.1109/ICIP.2016.7533003
25. 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: European Conference on Computer Vision, pp. 1–21 (2022). Springer
26. Aharon, N., Orfaig, R., Bobrovsky, B.-Z.: Bot-sort: Robust associations multi-pedestrian tracking, (2022). arXiv:2206.14651
27. Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9686–9696. (2023)
28. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: European Conference on Computer Vision, pp. 659–675 (2022). Springer
29. Zhang, Y., Wang, T., Zhang, X.: Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22056–22065 (2023)
30. Gao, R., Wang, L.: Memotr: Long-term memory-augmented transformer for multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9901–9910. (2023)

31. Theiner, J., Ewerth, R.: Tvcalib: Camera calibration for sports field registration in soccer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1166–1175. (2023)
32. Gutiérrez-Pérez, M., Agudo, A.: No bells just whistles: Sports field registration by leveraging geometric properties. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3325–3334. (2024)
33. Oo, Y.M., Jamsrandorj, A., Chao, V., Mun, K.-R., Kim, J.: A residual attention-based efficientnet homography estimation model for sports field registration. In: IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society, pp. 1–7 (2023). IEEE
34. Magera, F., Hoyoux, T., Barnich, O., Van Droogenbroeck, M.: Broadtrack: Broadcast camera tracking for soccer. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 6177–6187. IEEE (2025)
35. Magera, F., Hoyoux, T., Castin, M., Barnich, O., Cioppa, A., Van Droogenbroeck, M.: Can geometry save central views for sports field registration? In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 6070–6079. (2025)
36. Chan, A., Levine, M.D., Javan, M.: Player identification in hockey broadcast videos. Expert Syst. Appl. **165**, 113891 (2021). https://doi.org/10.1016/j.eswa.2020.113891
37. Balaji, B., Bright, J., Prakash, H., Chen, Y., Clausi, D.A., Zelek, J.: Jersey number recognition using keyframe identification from low-resolution broadcast videos. In: Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports, pp. 123–130. (2023)
38. Vats, K., McNally, W., Walters, P., Clausi, D.A., Zelek, J.S.: Ice hockey player identification via transformers and weakly supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3451–3460 (2022)
39. Koshkina, M., Elder, J.H.: A general framework for jersey number recognition in sports video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3235–3244. (2024)
40. Meta, A.: Llama 3.2: Revolutionizing edge ai and vision with open. Technical report, customizable models, vol. 9, Meta AI (2024). (Technical report)
41. Magera, F., Hoyoux, T., Barnich, O., Van Droogenbroeck, M.: A universal protocol to benchmark camera calibration for sports. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3335–3346. (2024)
42. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge university press, ??? (2003)
43. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: on the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vision **129**(11), 3069–3087 (2021)

Yewon Hwang is currently a postdoctoral researcher at the Korea Institute of Science and Technology (KIST). She received the Ph.D. in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea in 2024. Her current research interests include computer vision, multimodal learning, and natural language processing.

Yin May Oo is currently a Ph.D. candidate in AI-Robotics at the University of Science and Technology (UST), KIST School, in Seoul, South Korea. Her research interests primarily focus on computer vision, multimodal AI, sports video understanding, and sports analytics.

Muhammad Amrulloh Robbani is a Ph.D. candidate at the University of Science and Technology and a student researcher at the Korea Institute of Science and Technology, Seoul, South Korea. His research centers on multimodal computer vision for sport, large language models, and biomedical engineering for athletic performance. Motivated by an interest in sport and esports, he aims to develop intelligent, adaptive models grounded in domain-specific sport knowledge. He also has engineering experience in full-stack web and Android development and biomedical wearable devices.

Vanyi Chao is currently pursuing his Ph.D. degree at the Korea Institute of Science and Technology (KIST) and the University of Science and Technology (UST), Seoul, South Korea. His research interests include computer vision, deep learning, and AI-driven sports analysis. He aims to develop robust automated video analysis frameworks that extract fine-grained motion insights to evaluate player performance.

Ankhzaya Jamsrandorj is a Postdoctoral Researcher with the Intelligence and Interaction Research Center at the Korea Institute of Science and Technology (KIST), Seoul, South Korea. She received her Ph.D. degree in Computer Science from the KIST School, University of Science and Technology, in 2024. Her research interests include multimodal sports AI, digital biomarkers, vision-based gait analysis, and wearable sensor data analysis.

Hoang Quoc Nguyen is a researcher specializing in computer vision, deep learning, and video understanding. He graduated from the AI-Robotics program at the KIST School, University of Science and Technology (UST), in South Korea. His research interests include sports analytics, spatiotemporal event detection, human activity understanding, and multimodal AI.

Kyung-Ryoul Mun is a Principal Research Scientist at the Korea Institute of Science and Technology (KIST) and a Professor at the University of Science and Technology (UST). He received his Ph.D. in Biomedical Engineering from the National University of Singapore (NUS) in 2016. His research interests focus on Human Data Intelligence, including AI-driven healthcare, biomedical informatics, and digital biomarker discovery.

Jinwook Kim is currently a Principal Research Scientist at the Korea Institute of Science and Technology (KIST). He received his Ph.D. in Mechanical and Aerospace Engineering from Seoul National University in 2002. His research interests include human motion analysis, deep learning-based sports analytics, physics-based simulation, computer graphics, and healthcare technologies using multi-modal data.