Broadcast2Pitch: Game State Reconstruction from Unconstrained Soccer Videos

Yin May Oo[1,2*] Yewon Hwang[2*] Muhammad Amrulloh Robbani[1,2] Vanyi Chao[1,2]
Ankhzaya Jamsrandorj[2] Hoang Quoc Nguyen[2] Kyung-Ryoul Mun[2] Jinwook Kim[2]
[1]AI-Robotics, KIST School, University of Science and Technology, Seoul 02792, Republic of Korea
[2]Korea Institute of Science and Technology, Seoul 02792, Republic of Korea
{yinmay, ywhwang, amrullohrobbani, vanyi, ankhzaya, krmoon02}@kist.re.kr
nqhoang@ust.ac.kr, jinwook.kim21@gmail.com

Abstract

Game State Reconstruction (GSR) aims to reconstruct the 2D positions and identities of all athletes from broadcast soccer videos, requiring robust tracking, localization, and identity association under dynamic and unconstrained camera motions. We propose a modular GSR framework that integrates a multi-task keypoint and line detection model with an optimization-based homography estimation module. This approach leverages dense geometric cues from lines, circles, and keypoints to achieve robust spatial localization on a frame-by-frame basis, providing reliable alignment in diverse broadcast scenarios. To address identity consistency, we use appearance-based re-identification and a vision-language-guided tracklet refinement strategy to reduce ID switches and enforce temporal coherence. Comprehensive ablation studies validate the contribution of each component, and our framework achieves state-of-the-art performance on the SoccerNet-GSR benchmark, outperforming existing baselines by a significant margin. The proposed framework demonstrates strong robustness, generalization across scenes, and practical utility for structured game understanding in real-world broadcast sports analytics.

1. Introduction

Game State Reconstruction (GSR) is an emerging and impactful task in sports analytics that aims to produce a structured understanding of a match by localizing players on the pitch and identifying their jersey numbers, team affiliations, and tactical roles. Unlike standard multi-object tracking, GSR goes beyond trajectory estimation to recover a semantically rich and spatially grounded representation of the game. This holistic understanding unlocks a wide range of applications, including automated analytics, strategic evaluation, and immersive fan experiences. Soccer presents a particularly challenging testbed for this task due to its fast-paced dynamics and frequent visual ambiguities present in broadcast footage. Recent progress has been driven by the availability of large-scale public datasets and benchmarks [9–11, 14], which have catalyzed research in scalable, automated game understanding.

Achieving accurate GSR, however, remains highly challenging due to several factors: frequent occlusions, motion blur, varying camera viewpoints, and visually similar player appearances (*e.g.* identical jerseys across teammates), all of which complicate consistent tracking and identity resolution. Addressing these challenges requires the careful integration of multiple vision components, including multi-object tracking [3, 10, 12, 19, 40, 46, 50], player re-identification (ReID), sports field registration [6, 16, 22, 32, 33, 37], and fine-grained athlete identity recognition [2, 5, 24, 44, 45]. The GSR Baseline [39] offers a strong starting point, but it struggles with spatial misalignment and frequent identity switches, particularly under visually ambiguous or occluded conditions. These limitations motivate the need for a more robust and modular framework capable of jointly addressing localization, tracking, and identity reasoning in broadcast sports video.

To address these challenges, we propose a modular framework that achieves state-of-the-art results on the SoccerNet-GSR benchmark [39]. Our system follows a tracking-by-detection paradigm, using YOLOX [13] for player detection, DeepEIoU [20] for trajectory linking, and OSNet [52] for appearance-based re-identification. To enhance identity consistency across tracklets, we introduce a split-and-merge tracklet refinement strategy guided by identity cues from a single vision-language

*Equal contribution.

model (VLM). In contrast to prior approaches [15, 39] that rely on separate classifiers for role, team, and jersey number, we formulate identity recognition as a visual question answering (VQA) task. A generative VLM [1, 30, 47] is prompted with natural language queries over cropped player images to jointly predict multiple identity attributes. For sports field registration, we propose a homography estimation approach based on a unified optimization that integrates dense geometric cues including lines, circles, and keypoints.

The main contributions of this paper are as follows:

1. A state-of-the-art modular framework for GSR that integrates detection, tracking, identity recognition, and sports field registration into a unified pipeline, enabling robust analysis of broadcast soccer videos.
2. Achieving state-of-the-art performance in sports field registration on the SoccerNet-GSR benchmark, our joint optimization effectively leverages keypoints, lines, and circle constraints.
3. A tracklet refinement strategy leveraging VLM-predicted identity cues to reduce ID switches and enhance temporal coherence of tracklets.

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*. [39] introduced the first end-to-end minimap reconstruction pipeline under the SoccerNet benchmark [7–9, 11, 14, 31], combining detection [4, 13, 35, 36, 51], tracking, re-identification [17, 29, 38, 52, 53], and sports field registration to provide structured game state outputs. Building upon this, Golovkin *et al*. [15] proposed improvements in camera calibration, spatial alignment, and identity continuity, achieving state-of-the-art performance on the GSR task.

2.2. Sports Field Registration

Sports field registration is essential for projecting player positions onto the pitch under challenging broadcast views. TVCalib [42] performs 3D calibration using line and circle reprojection error without keypoints. NBJW [16] 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 [27] integrates optical flow with detailed camera models for smooth tracking, while Magera *et al*. [28] 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.

2.3. Athlete Identification

Consistent identity tracking remains a core challenge in GSR. Traditional appearance-based ReID methods [52, 53] often struggle under occlusions and uniform similarity. Jersey number recognition offers a more robust alternative. Early work by Chan *et al*. [5] combined CNN-LSTM architectures, while Balaji *et al*. [2] selected keyframes with clear jersey visibility. Recent approaches employ transformers [24, 44] or vision-language models [17] under weak supervision. PRTreID [29] learns foreground embeddings from body parts and jointly predicts player role and team affiliation through multi-task learning. Other recent approaches [15, 39] 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 [13] for detection, DeepEIoU [20] for tracking, which utilizes OSNet [52] ReID features with expansion IoU for improved association. LLaMA-3.2-Vision [30] 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

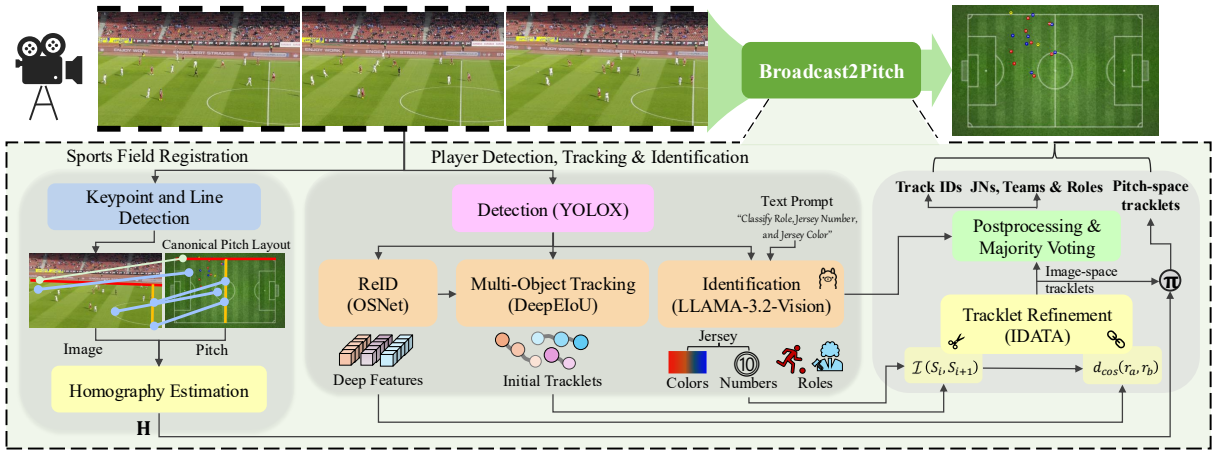

Figure 1. Overview of our Broadcast2Pitch framework.

position on the ground plane. This mapping is modeled by a homography matrix $\mathbf{H} \in \mathbb{R}^{3\times3}$, which transforms image-space coordinates $\mathbf{p}_i \in \mathbb{R}^2$ into pitch coordinates $\mathbf{P}_i \in \mathbb{R}^2$:

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

where $\tilde{\mathbf{p}}_i = [x_i, y_i, 1]^\top$ 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 $\mathbf{H}$ using a unified optimization that minimizes the residuals between a canonical pitch template of standardized dimensions (*e.g.*, 105 m $\times$ 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).

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. 2(a)) and 18 pitch lines (Fig. 2(b)) in [26], 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 $\mathbf{M}^{\text{kpt}} \in [0,1]^{97\times H\times W}$ and line heatmaps $\mathbf{M}^{\text{line}} \in [0,1]^{18\times H\times W}$. 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 $\mathbf{G} = \mathbf{H}^{-1}$, 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. 2(b)), we extract high-confidence samples $\{(x_{L,j}, y_{L,j}, w_{L,j})\}_{j=1}^{N}$ from its predicted heatmap, $\mathbf{M}^{\text{line}}$ 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 $\ell'_L \sim \mathbf{G}^\top \ell_L$. For each image-space sample $\tilde{\mathbf{p}}_{L,j} = [x_{L,j}, y_{L,j}, 1]^\top$, we compute a confidence-weighted algebraic line residual:

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

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

Circle (conic) constraint. Let $\mathbf{C}$ denote the center circle on the pitch, centered at $\mathbf{o} = (x_0, y_0, 1)^\top$ with radius r (e.g., 9.15m). In the pitch template space, this circle is

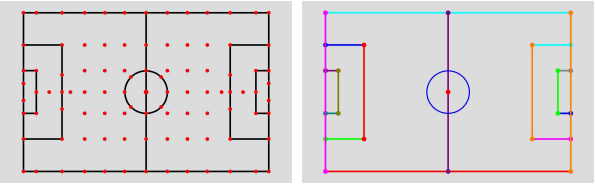

Figure 2. **Canonical pitch layout with keypoint and line representations.** (a) The 97 pitch keypoints (red dots) serve as discrete anchors for localization. (b) The 18 pitch lines (colored polylines) capture extended geometric structures, including sidelines, penalty areas, and the center circle. Together,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

represented as a conic, defined by the matrix:

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

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

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

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

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

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

Homography objective and solver. We estimate $\mathbf{G}$ by minimizing the least-squares objective of lines, circle, and keypoints:

$$\min_{\mathbf{G}} \underbrace{\sum_L \sum_j \left(r_{L,j}^{\text{line}}(\mathbf{G})\right)^2}_{\text{Lines}} + \underbrace{\sum_k \left(r_k^{\text{circ}}(\mathbf{G})\right)^2}_{\text{Circle}} + \underbrace{\sum_i \left(r_i^{\text{kpt}}(\mathbf{G})\right)^2}_{\text{Keypoints}} \tag{6}$$

We initialize $\mathbf{G}$ using a DLT[18] 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. 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 [20, 46]. We formulate identity recognition (jersey number, color, and role) as a multimodal autoregressive generation task. Specifically, we leverage LLaMA-3.2-Vision [30], 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:

$$P(x \mid q, v) = \prod_{t=1}^{T} P(x_t \mid x_{<t}, q, v) \tag{7}$$

where $x = (x_1, x_2, \ldots, x_T)$ denotes the sequence of identity tokens, q is the instruction prompt (*e.g.*, "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.

3.3. Tracklet Refinement

After generating initial tracklets with DeepEIoU [20], we apply a split-and-merge refinement strategy inspired by [41], with a key distinction: rather than relying on appearance features (*e.g.*, ReID) for splitting, we leverage frame-wise identity predictions (jersey number, color, role) obtained from the VLM, hence the name ID-Aware Tracklet Refinement (IDATR). This avoids unnecessary splits caused by unreliable ReID features under occlusion—common in sports footage. For merging, we incorporate identity consistency in addition to the spatial-temporal constraints of [41]. Below, we present the formulation of our split-and-merge method.

Split. We observe that identity switches often arise when a tracklet contains a large temporal gap, typically

reflecting a player's exit and re-entry. In such cases, appearance- and IoU-based methods [20] often fail due to similar visual features across players or unreliable IoU due to large frame jumps. 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 split points as $G = \{g \mid \Delta_f > \tau_{gap}\}$, and segment the tracklet into $[S_1, S_2, ..., S_{|G|+1}]$. We then evaluate whether adjacent segments S_i and S_{i+1} should remain split by checking consistency in jersey number, role, and color, using an identity inconsistency indicator, $\mathcal{I}$:

$$\mathcal{I}(X,Y) = \begin{cases} & \text{if Maj}(R_X) \neq \text{Maj}(R_Y) \\ 1 & \vee\ \text{Maj}(J_X) \neq \text{Maj}(J_Y) \\ & \vee\ \text{Maj}(C_X) \neq \text{Maj}(C_Y) \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where $\text{Maj}(R_X)$, $\text{Maj}(J_X)$, and $\text{Maj}(C_X)$ denote the majority-voted role, jersey number, and color for X, respectively. If $\mathcal{I}(S_i, S_{i+1})$ is equal to 1, then split occurs.

Merge. After splitting, we perform a merging step to produce coherent tracklets. Two tracklets T_a and T_b are merged if their ReID features r_a and r_b are sufficiently similar, as measured by cosine distance:

$$d_{cos}(r_a, r_b) = 1 - \frac{r_a \cdot r_b}{\|r_a\|\|r_b\|} < \tau_{cos}. \tag{9}$$

To ensure robustness, we incorporate the following constraints: (1) *Spatial constraint* — $\|p_a^{end} - p_b^{start}\|_2 < \tau_{spatial}$, where p_a^{end} and p_b^{start} denote the spatial positions at the end of T_a and the start of T_b, respectively, and $\tau_{spatial}$ is a spatial continuity threshold. (2) *Temporal constraint* — $F_a \cap F_b = \emptyset$, where F_a and F_b are the sets of frame indices occupied by T_a and T_b, ensuring the same identity does not appear in multiple locations at the same time. (3) *Identity consistency constraint* — $\mathcal{I}(T_a, T_b) = 0$. Formally, tracklet merge occurs when $\mathcal{M}(T_a, T_b) = 1$, where

$$\mathcal{M}(T_a,T_b) = \begin{cases} 1, & \text{if } d_{cos}(r_a, r_b) < \tau_{cos} \\ & \quad \wedge\ \|p_a^{end} - p_b^{start}\|_2 < \tau_{spatial} \\ & \quad \wedge\ F_a \cap F_b = \emptyset \\ & \quad \wedge\ \mathcal{I}(T_a, T_b) = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

3.4. Post-processing

After obtaining tracklets from IDATR and LLaMA-3.2-Vision [30], 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. We further discuss the validity of our assumption in Section 4.6. 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 [39], which contains 200 broadcast video clips, each 30 seconds long. It extends SoccerNet-Tracking [9] 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 [39], a metric specifically designed to assess GSR by jointly measuring detection accuracy, identity association, and localization accuracy. Formally, GS-HOTA is computed as follows:

$$GS{-}HOTA(P,G) = LocSim(P,G) \times IdSim(P,G), \tag{11}$$

where

$$LocSim(P,G) = e^{\ln(0.05)\frac{||P-G||_2^2}{r^2}}, \tag{12}$$

and

$$IdSim(P,G) = \begin{cases} 1 & \text{if all ID attributes match,} \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

$LocSim(P,G)$ 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

Framework	Detection	Tracking	Sport Field Registration	Athlete Identification
GSR Baseline [39]	YOLOv8 [43]	StrongSORT [12]	TVCalib [42]	PRTReID [29] + MMOCR [25]
Golovkin *et al.* [15]	YOLOv5 [23]	DeepSORT [46]	SegFormer [48] + ResNet50	OSNet [52] + ResNet-18 + WideResNet [49]
Ours	YOLOX [13]	DeepEIoU [20]	EfficientNet-Attention U-Net [33]	LLaMA-3.2-Vision [30]

Table 1. **Component-level comparison of Game State Reconstruction (GSR) frameworks.** The table highlights the backbone models chosen for each core module—detection, tracking, field registration, and athlete identification—showing the differences between prior works and our approach.

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.

Sports Field Registration. We evaluate homography accuracy using standard calibration metrics [26]. The Jaccard index for calibration (JaC) measures the percentage of field markings correctly reprojected within a pixel threshold; we report JaC_5 and JaC_{10} at 5 and 10 pixels, respectively. Mean Reprojection Error (MRE) and Median Reprojection Error (MedRE) measure the average and median pixel distance between the projected field markings and the ground truth. Completeness Rate (CR) denotes the percentage of frames with a valid homography estimation.

4.3. Implementation Details

We employ YOLOX [13], trained on the SoccerNet-Tracking dataset [9], and OSNet [52], trained on the SportsMOT dataset [10], 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 [39] train set, each resized to 384×384 pixels. The LLaMA-3.2-Vision model is fine-tuned [30] on the SoccerNet jersey dataset [8] for just one epoch with a learning rate of $1e^{-5}$. For the hyperparameters used in IDATR, we set $\tau_{\text{gap}} = 25$, $\tau_{\text{cos}} = 0.4$, and $\tau_{\text{spatial}} = 0.8$. All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24 GB memory using the PyTorch framework.

4.4. Quantitative Evaluation

We evaluate the effectiveness of our Broadcast2Pitch framework as well as our homography model on the SoccerNet-GSR test set. This dual evaluation provides a comprehensive assessment of player localization, identity tracking, and geometric alignment of the pitch model.

Framework	GS-HOTA ↑	GS-DetA ↑	GS-AssA ↑	IDF1 ↑
GSR Baseline [39]	22.26	10.67	46.46	-
Golovkin *et al.* [15]	55.82	41.67	74.86	-
Ours	**61.48**	**48.47**	**78.00**	**64.20**

Table 2. **Performance comparison on the SoccerNet-GSR test set.** We compare our framework with other baselines on GS-HOTA and related metrics.

GSR. Table 2 presents the quantitative results on the SoccerNet-GSR test set, comparing our method against the GSR Baseline [39] and Golovkin *et al.* [15]. The corresponding framework details are provided in Table 1. Our approach achieves a GS-HOTA score of 61.48, substantially outperforming Golovkin *et al.* and far surpassing the original GSR baseline. We also obtain the highest reported scores in detection accuracy (GS-DetA: 48.47) and association accuracy (GS-AssA: 78.00), indicating both precise spatial localization and robust temporal association. In addition, our method attains an IDF1 score of 64.20, demonstrating strong temporal consistency in player tracking. These results establish a new state-of-the-art on SoccerNet-GSR test set and highlight the effectiveness of our integrated framework, combining vision-language identity reasoning, homography estimation, and tracklet refinement under unconstrained broadcast video conditions.

Sports Field Registration. For the task of sports field registration, our method establishes a new state-of-the-art on the SoccerNet-GSR test set, with results detailed in Table 3. Our approach significantly outperforms prior works, achieving a JaC_5 of 69.38 and a JaC_10 of 92.84. This result demonstrates robust performance at the strict 5-pixel threshold and comprehensive accuracy at the 10-pixel threshold. We also report the lowest reprojection errors (MRE: 4.47, MedRE: 2.21), further highlighting the precision of our method. We attribute the significant performance gains to the proposed unified homography estimation, which integrates line, circle, and fallback keypoint constraints within a single optimization framework.

4.5. Qualitative Evaluation

Figure 3 presents qualitative comparisons with the GSR baseline [39]. While both methods detect athletes across

Method	JaC_5 ↑	JaC_{10} ↑	MRE ↓	MedRE ↓	CR ↑
TVCalib [42]	19.88	50.42	12.40	–	99.93
NBJW [16]	37.14	68.24	10.28	–	93.67
BroadTrack [27]	56.88	79.79	5.02	2.37	100
Oo *et al.* * [33]	40.44	68.34	9.36	3.70	99.98
Ours*	**69.38**	**92.84**	**4.47**	**2.21**	**100**

Table 3. **Evaluation of sports field registration performance on SoccerNet-GSR test set.** Metrics for TVCalib, NBJW, and BroadTrack are reported from [27] while methods with an asterisk (*) are evaluated with our implementation from [26].

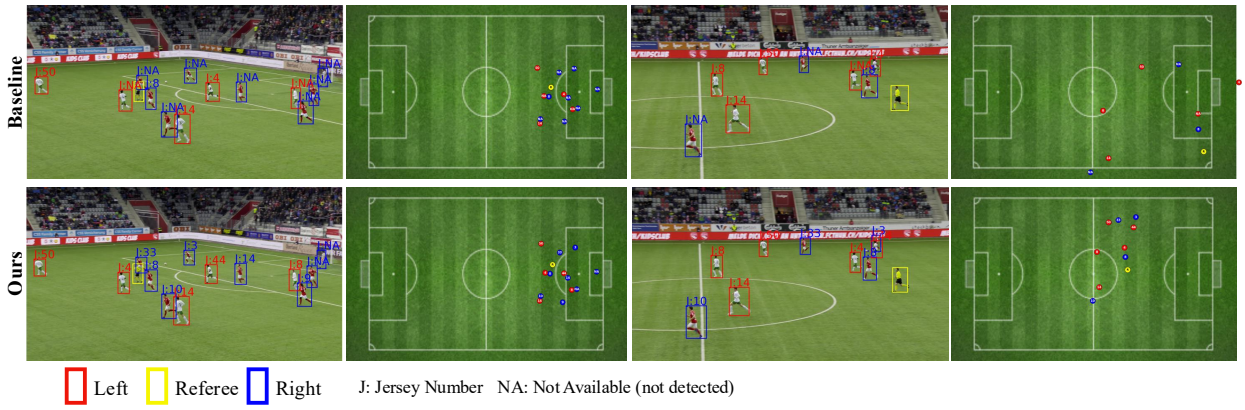

Figure 3. Qualitative comparison on SoccerNet-GSR test set. Our results (bottom) show improved jersey number recognition and pitch localization compared to the baseline (top).

Modules			Metrics			
HE(Keypoint)	HE(Line)	IDATR	GS-HOTA ↑	GS-DetA ↑	GS-AssA ↑	IDF1 ↑
✓	×	×	48.23	34.77	66.91	51.68
×	✓	×	56.39	42.77	74.40	61.31
✓	✓	×	58.51	44.63	76.72	61.57
✓	✓	✓	**61.48**	**48.47**	**78.00**	**64.20**

Table 4. **Ablation study on homography estimation (HE) and tracklet refinement.** The detector (Sec. 3.1.1) is identical across all settings. We vary the HE module (Sec. 3.1.2) by using only keypoints, only lines, or both. IDATR denotes the Identity-Aware Tracklet Refinement module.

frames, our approach achieves notably more accurate pitch localization through robust homography estimation and significantly improves jersey number recognition – even under severe occlusions compared to the baseline. Additional qualitative examples over full sequences are provided in the supplementary videos. Table 6 illustrates the impact of IDATR on identity consistency. In the first row, the same player is incorrectly split into two tracklets, while in the second row, two distinct players are mistakenly merged into one. Applying IDATR effectively resolves both issues, yielding more coherent and accurate player tracklets over time.

4.6. Ablation Study

Sports Field Registration and IDATR. We perform ablation studies to assess the contribution of homography estimation and IDATR (Table 4). We keep the keypoint and line detection model (Sec. 3.1.1) fixed and vary the geometric cues used during homography estimation. Using only keypoint predictions with DLT and RANSAC yields a GS-HOTA of 48.23, reflecting the limitations of sparse correspondences under broadcast conditions. Switching to our optimization using line and circle constraints (Sec. 3.1.2), without keypoint fallback, significantly improves GS-HOTA to 56.39, highlighting the strength of dense geometric constraints. Integrating lines, circles, and keypoints in the joint optimization further improves alignment, reaching 58.51 GS-HOTA. Finally, adding our IDATR boosts performance to 61.48 GS-HOTA and 64.20 IDF1, underscoring the synergy between accurate spatial registration and consistent identity association.

Athlete Identification and GS-HOTA Components. To select the best athlete identification model for our Broadcast2Pitch, we compare three different approaches, with results shown in Table 5. The first approach employs specialized models tailored to each identity attribute: PRTReID [29] for role classification, EasyOCR [21] for jersey number recognition, and ResNet-18 trained on the SoccerNet jersey dataset [8] for jersey color classification, which is subsequently used for team affiliation. The second approach explores the use of a vision-language model (VLM) by leveraging a CLIP [34] encoder for general-purpose representation learning, combined with attribute-specific classification heads for role, jersey number, and team prediction. The third approach corresponds to our proposed method. The results show that VLMs are significantly more effective than using separate, attribute-specific models. Notably, CLIP achieves strong results (GS-HOTA 60.13, IDF1 62.20), demonstrating the benefits of large-scale pretraining. Our proposed LLaMA-3.2-Vision achieves the best overall performance (GS-HOTA 61.48, IDF1 64.20) by jointly reasoning over role, team, and jersey attributes in a unified backbone. We adopt LLaMA-3.2-Vision for its coherent multi-attribute reasoning and simplified architecture, while acknowledging that its modest accuracy gain over CLIP comes at higher computational cost.

To assess how incorporating identity attributes into the evaluation metric impacts overall performance, we conduct an ablation study, with results summarized in Table 7. Adding identity attributes to the evaluation metric (rows 2–6) imposes stricter criteria compared to the pitch-only setting (row 1), which solely measures spatial tracking accuracy. In these stricter settings, the GS-

Model	GS-HOTA ↑	GS-DetA ↑	GS-AssA ↑	IDF1 ↑
PRTReID [29] + EasyOCR [21] + ResNet-18	18.11	6.54	50.14	10.62
CLIP [34]	60.13	46.88	77.15	62.20
LLaMA-3.2-Vision [30] (Ours)	**61.48**	**48.47**	**78.00**	**64.20**

Table 5. **Comparison of athlete identification models on athlete identification within our GSR framework.** LLaMA-3.2-Vision achieves the best performance on the SoccerNet-GSR test set.

HOTA score is penalized even when the spatial localization is correct if any of the specified identity attributes are misclassified. Accordingly, we treat the pitch-only setting as an upper bound and examine the performance degradation as additional attributes are introduced.

Incorporating role or team into the evaluation results in only a marginal drop in GS-HOTA, not only indicating that the VLM handles these attributes effectively, but also supporting the validity of our underlying assumption for team affiliation. However, the inclusion of jersey numbers leads to a substantial decline in performance, identifying jersey recognition as the primary bottleneck. This difficulty arises from the nature of broadcast footage: jersey digits are often blurred, occluded, or visible in only a few frames of a tracklet. The challenge is further exacerbated by the tracklet-level evaluation protocol: jersey number predictions are aggregated via majority voting, which is particularly fragile when correct digits are sparsely observed. Further, although IDATR enhances initial tracking performance by improving identity association, it does not produce perfect tracklets. As a result, residual ID switches can still occur post-IDATR, leading to erroneous jersey number predictions and further degrading the final GS-HOTA score.

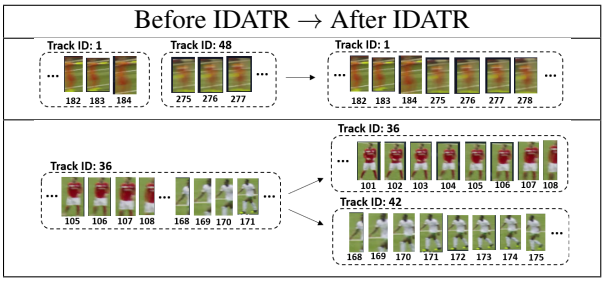

Table 6. **Comparison of tracking results before and after IDATR.**

5. Conclusion

We propose a modular and robust framework for Game State Reconstruction (GSR) from unconstrained broadcast soccer videos, addressing key challenges in player localization, spatial alignment, and identity preservation. Our method is built on three core components: (1) a multi-task model that detects keypoints and lines, providing dense geometric cues for a unified, optimization-based homography estimation, (2) a vision-language model that performs joint recognition of player roles, teams, and jersey numbers from a single input, and (3) IDATR – a tracklet refinement strategy that leverages vision-language identity cues to resolve association errors and ensure temporal coherence. Evaluated on the SoccerNet-GSR benchmark with the GS-HOTA metric suite, our approach achieves state-of-the-art performance across all evaluation metrics. Ablation studies confirm the effectiveness of each module, showing that leveraging dense lines, circle and keypoints constraints significantly improves homography estimation over using keypoints alone, while identity-aware tracklet association offer complementary improvements in both accuracy and consistency.

GS-HOTA Components				Metrics			
Pitch	Role	Team	Jersey	GS-HOTA ↑	GS-DetA ↑	GS-AssA ↑	IDF1 ↑
✓	×	×	×	79.52	85.97	73.60	83.40
✓	✓	×	×	79.20	85.18	73.68	83.14
✓	×	✓	×	73.58	74.46	72.77	76.97
✓	×	✓*	×	77.39	81.35	73.66	81.17
✓	×	×	✓	64.70	53.96	77.59	68.97
✓	✓	✓	✓	61.48	48.47	78.00	64.20

Table 7. **Ablation study on GS-HOTA components.** We evaluate the impact of incorporating Role, Team, and Jersey attributes—alongside Pitch—on overall performance. Samples SNGS-126, 131, 197 are incorrectly annotated for the team attribute in the SoccerNet-GSR test set. Asterisk (*) indicates the exclusion of these samples.

6. Limitations and Future Work

While our ablation study demonstrates that the assumption used for team affiliation is generally robust, we acknowledge that it is not universally reliable. The assumption may break down in edge cases such as corner kicks or sustained high-pressing scenarios. Additionally, the LLaMA-3.2-Vision model achieves state-of-the-art results in identity recognition by jointly reasoning over multiple attributes. However, its high computational cost presents a significant challenge for real-time deployment. Although IDATR improves association accuracy, it requires task-specific hyperparameter tuning and limits its generalizability across different domains. As future work, we aim to make the tracklet refinement module learnable, enabling it to adapt to diverse scenarios and generalize beyond the specific conditions of the initial tracking module. This would allow the refinement process to move beyond fixed heuristics and better handle variations in player motion, occlusion, and identity switches. Overall, our findings highlight promising directions for joint vision-language reasoning and adaptive identity modeling, particularly in complex, real-world sports analytics scenarios.

7. Acknowledgement

This work was supported by Development of foot diagnosis and shoe manufacturing technology for the disabled Project (RS-2022-00164554) and the Korea Institute of Science and Technology (KIST) Institutional Program (No. 2E33841).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[2] Bavesh Balaji, Jerrin Bright, Harish Prakash, Yuhao Chen, David A Clausi, and John Zelek. Jersey number recognition using keyframe identification from low-resolution broadcast videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 123–130, 2023. 1, 2

[3] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9686–9696, 2023. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[5] Alvin Chan, Martin D. Levine, and Mehrsan Javan. Player identification in hockey broadcast videos. *Expert Systems with Applications*, 165:113891, 2021. 1, 2

[6] Yen-Jui Chu, Jheng-Wei Su, Kai-Wen Hsiao, Chi-Yu Lien, Shu-Ho Fan, Min-Chun Hu, Ruen-Rone Lee, Chih-Yuan Yao, and Hung-Kuo Chu. Sports field registration via keypoints-aware label condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3523–3530, 2022. 1

[7] Anthony Cioppa, Adrien Deliege, Floriane Magera, Silvio Giancola, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. 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*, pages 4537–4546, 2021. 2

[8] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up soccernet with multi-view spatial localization and re-identification. *Scientific Data*, 9, 2022. 6, 7

[9] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3491–3502, 2022. 1, 2, 5, 6

[10] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9921–9931, 2023. 1, 6

[11] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. 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*, pages 4508–4519, 2021. 1, 2

[12] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 25:8725–8737, 2023. 1, 6

[13] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1, 2, 6

[14] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1711–1721, 2018. 1, 2

[15] Vladimir Golovkin, Nikolay Nemtsev, Vasyl Shandyba, Oleg Udin, Nikita Kasatkin, Pavel Kononov, Anton Afanasiev, Sergey Ulasen, and Andrei Boiarov. From broadcast to minimap: Achieving state-of-the-art soccernet game state reconstruction. *arXiv preprint arXiv:2504.06357*, 2025. 2, 6

[16] Marc Gutiérrez-Pérez and Antonio Agudo. No bells just whistles: Sports field registration by leveraging geometric properties. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3325–3334, 2024. 1, 2, 6

[17] Konrad Habel, Fabian Deuser, and Norbert Oswald. Clip-reident: Contrastive training for player re-identification. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, pages 129–135, 2022. 2

[18] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[19] Hsiang-Wei Huang, Cheng-Yen Yang, Samartha Ramkumar, Chung-I Huang, Jenq-Neng Hwang, Pyong-Kun Kim, Kyoungoh Lee, and Kwangju Kim. Observation centric and central distance recovery for athlete tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 454–460, 2023. 1

[20] Hsiang-Wei Huang, Cheng-Yen Yang, Jiacheng Sun, Pyong-Kun Kim, Kwang-Ju Kim, Kyoungoh Lee, Chung-I Huang, and Jenq-Neng Hwang. 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, pages 163–172, 2024. 1, 2, 4, 5, 6

[21] JaidedAI. Easyocr. https://github.com/JaidedAI/EasyOCR, 2019. 7, 8

[22] Wei Jiang, Juan Higuera, Baptiste Angles, Weiwei Sun, Mehrsan Javan Roshtkhari, and Kwang Yi. Optimizing through learned errors for accurate sports field registration. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 201–210, 2020. 1

[23] Glenn Jocher. Ultralytics yolov5, 2020. 6

[24] Maria Koshkina and James H Elder. A general framework for jersey number recognition in sports video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3235–3244, 2024. 1, 2

[25] Zhanghui Kuang, Hongbin Sun, Zhizhong Li, Xiaoyu Yue, Tsui Hin Lin, Jianyong Chen, Huaqiang Wei, Yiqin Zhu, Tong Gao, Wenwei Zhang, et al. Mmocr: a comprehensive toolbox for text detection, recognition and understanding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3791–3794, 2021. 6

[26] Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. A universal protocol to benchmark camera calibration for sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3335–3346, 2024. 3, 6

[27] Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. Broadtrack: Broadcast camera tracking for soccer. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6177–6187. IEEE, 2025. 2, 6

[28] Floriane Magera, Thomas Hoyoux, Martin Castin, Olivier Barnich, Anthony Cioppa, and Marc Van Droogenbroeck. Can geometry save central views for sports field registration? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6070–6079, 2025. 2

[29] Amir M Mansourian, Vladimir Somers, Christophe De Vleeschouwer, and Shohreh Kasaei. 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*, pages 103–112, 2023. 2, 6, 7, 8

[30] AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open. Technical report, customizable models. Technical report, Meta AI, 9 2024. 2, 4, 5, 6, 8

[31] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5085, 2023. 2

[32] Xiaohan Nie, Shixing Chen, and Raffay Hamid. A robust and efficient framework for sports-field registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1936–1944, 2021. 1

[33] Yin May Oo, Ankhzaya Jamsrandorj, Vanyi Chao, Kyung-Ryoul Mun, and Jinwook Kim. A residual attention-based efficientnet homography estimation model for sports field registration. In *IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–7. IEEE, 2023. 1, 2, 3, 6

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7, 8

[35] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023. 2

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[37] Feng Shi, Paul Marchwica, Juan Camilo Gamboa Higuera, Mike Jamieson, Mehrsan Javan, and Parthipan Siva. Self-supervised shape alignment for sports field registration. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3768–3777, 2022. 1

[38] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1613–1623, 2023. 2

[39] Vladimir Somers, Victor Joos, Anthony Cioppa, Silvio Giancola, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir M Mansourian, Xin Zhou, Shohreh Kasaei, et al. 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*, pages 3293–3305, 2024. 1, 2, 5, 6

[40] Vladimir Somers, Baptiste Standaert, Victor Joos, Alexandre Alahi, and Christophe De Vleeschouwer. Cameltrack: Context-aware multi-cue exploitation for online multi-object tracking. *arXiv preprint arXiv:2505.01257*, 2025. 1

[41] Jiacheng Sun, Hsiang-Wei Huang, Cheng-Yen Yang, Zhongyu Jiang, and Jenq-Neng Hwang. Gta: Global tracklet association for multi-object tracking in sports. In *Proceedings of the Asian Conference on Computer Vision*, pages 421–434, 2024. 4

[42] Jonas Theiner and Ralph Ewerth. Tvcalib: Camera calibration for sports field registration in soccer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1166–1175, 2023. 2, 6

[43] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024. 6

[44] Kanav Vats, William McNally, Pascale Walters, David A Clausi, and John S Zelek. Ice hockey player identification via transformers and weakly supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3451–3460, 2022. 1, 2

[45] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A Clausi, and John S Zelek. Player tracking and identification in ice hockey. *Expert systems with applications*, 213:119250, 2023. 1

[46] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1, 4, 6

[47] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 2

[48] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 6

[49] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6

[50] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 1

[51] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. 2

[52] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3702–3712, 2019. 1, 2, 6

[53] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5056–5069, 2021. 2